



Building a multimodal entity linking dataset from tweets

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, Brigitte Grau

► To cite this version:

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, Brigitte Grau. Building a multimodal entity linking dataset from tweets. LREC 2020 - Language Resources and Evaluation Conference, May 2020, Marseille, France. pp.4885-4292. hal-04315504

HAL Id: hal-04315504

<https://hal.science/hal-04315504>

Submitted on 30 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Building a Multimodal Entity Linking Dataset From Tweets

Omar Adjali¹, Romaric Besançon¹, Olivier Ferret¹, Hervé Le Borgne¹, Brigitte Grau²

1. CEA, LIST, F-91191 Gif-sur-Yvette, France

2. LIMSI, CNRS, F-91405 Orsay, France, Université Paris-Saclay
 {omar.adjali, romaric.besancon, olivier.ferret, herve.le-borgne}@cea.fr
 brigitte.grau@limsi.fr

Abstract

The task of Entity linking, which aims at associating an entity mention with a unique entity in a knowledge base (KB), is useful for advanced Information Extraction tasks such as relation extraction or event detection. Most of the studies that address this problem rely only on textual documents while an increasing number of sources are multimedia, in particular in the context of social media where messages are often illustrated with images. In this article, we address the Multimodal Entity Linking (MEL) task, and more particularly the problem of its evaluation. To this end, we propose a novel method to quasi-automatically build annotated datasets to evaluate methods on the MEL task. The method collects text and images to jointly build a corpus of tweets with ambiguous mentions along with a Twitter KB defining the entities. We release a new annotated dataset of Twitter posts associated with images. We study the key characteristics of the proposed dataset and evaluate the performance of several MEL approaches on it.

Keywords: Entity linking, social media, multimodality, multimedia entity linking

1. Introduction

The emergence of social media platforms allows gathering a lot of data where text and images are related. Yet, most of the Information Retrieval or Information Extraction tasks still mainly rely on textual data to automatically extract meaningful knowledge from unstructured text. In particular, over the last decade, the Twitter platform was a major source of information for natural language processing (NLP) applications such as information retrieval, sentiment analysis, or topic modeling. However, the analysis of social media posts poses new challenges to NLP researchers since, despite a large amount of available data, social media posts are often short and noisy, making information extraction tasks more difficult. Among these tasks, one key processing step is to map a named mention (also known as surface form) within a text to an actual entity defined in a knowledge base. This task is referred to as Entity Linking (EL) or named entity disambiguation (NED). Applications such as event extraction, knowledge base population, and relation extraction can directly benefit from Entity linking. Early approaches of EL (Bunescu and Paşca, 2006; Cucerzan, 2007; Milne and Witten, 2008; Dredze et al., 2010) exploit lexical entity mention similarities, entity mention context similarities and prior information about entity candidates in the KB to rank them, while global approaches (Ratinov et al., 2011; Guo and Barbosa, 2014; Pershina et al., 2015; Globerson et al., 2016) leverage all kinds of relational features between mentions within a document and between entities in a knowledge base to globally resolve them. The Entity linking task has been traditionally achieved on documents such as newspaper articles, which is rich in textual content but generally lacks visual information that could be used in the EL task. On the other hand, social media posts like tweets provide poor and noisy textual contexts which make the Entity linking task harder, but is often associated with complementary visual information.

In this paper, we propose to automatically build large-scale Multimodal Entity Linking (MEL) datasets and formulate

the MEL problem on tweets where textual and visual information are systematically exploitable. To avoid a tremendous manual annotation effort, we rely on a Twitter-specific mechanism that allows us to jointly create ambiguous mentions and generate their corresponding entity candidates. More specifically, we exploit the fact that Twitter users can make use of Twitter *screen names* (e.g., @UserScreenName) in their tweet posts to mention other users, which provides us with unambiguous mentions. However, we observe that many tweets also contain proper names (e.g. last names or acronyms for organizations) to refer to other Twitter users, thereby creating ambiguities about the user (entity) they refer to. Consider the following real-world example tweet:



This example shows how named mentions of users (here, *Bloomberg* and *Trump*) may be ambiguous with users that share the same last names when they are not explicitly represented using their corresponding Twitter screen names, respectively @MikeBloomberg and @realDonaldTrump. This example illustrates both the need to perform EL on tweets and the possibility to automatically create ambiguities and build a large scale MEL dataset for tweets, by replacing screen names by ambiguous mentions. Figure 1 illustrates example tweets from our dataset. Each row represents a tweet posted by a given user u_i who refers in its tweet to another user u_j . We note that each mention is characterized by a textual context (tweet) and its corresponding visual context (image). Moreover, as most Twitter users do not have entries in standard knowledge bases (e.g. Freebase, Wikidata), we propose to link ambiguous mentions to a specifically built Twitter user knowledge base. Figure 3 shows some example entities from our Twitter KB.

Our contributions are summarized as follows:

Textual Context	Visual Context	Ground truth	Mention
Only a few hours now until the start of the ground-breaking @CVF #VirtualClimateSummit. Very glad that gore will be part of it. His unwavering advocacy for ambitious climate action is inspirational.		@algore	gore
Former @Megadeth star friedman is making his way to the @thehiifindy to melt your hearts this Valentines Day		@marty_friedman	friedman
JUST ANNOUNCED: One of the worlds foremost advocates for women, gates shares her journey towards gender equity in THE MOMENT OF LIFT. Tickets on sale Wednesday at 10:00 am		@melindagates	gates

Figure 1: Example of samples of the proposed corpus. Considering a tweet illustrated with an image (*left*), an entity mention is created by replacing the original screen name (*middle*) by an appropriate text (*right*).

- we investigate the Multimodal Entity Linking (MEL) task and define textual and visual representations for both mentions and entities;
- we elaborate a method for automatically building MEL datasets based on tweets. We provide a new dataset based on this method and present an analysis of its main features;
- we propose a first model that achieves interesting results on the proposed dataset, validating our approach and highlighting the potential performance gains of integrating textual and visual information in more complex models.

The dataset, the code to produce a new one and to evaluate with the methods described in this article will be released at https://github.com/OA256864/MEL_Tweets

2. Related Work

The MEL task is intrinsically related to several research domains: first, EL from text, which is characterized by a set of evaluation frameworks. But EL was also applied to specific forms of text such as social media posts. Finally, MEL is closely related to work about multimedia as it combines text and image contents. Since we propose a new corpus, we particularly focus on the corpora developed in these three areas.

Entity linking corpora. Existing corpora for evaluating entity linking systems are thoroughly reviewed in (Ling et al., 2015; Usbeck et al., 2015; Van Erp et al., 2016; Rosales-Méndez, 2019). They were mostly constructed from news articles: AIDA-YAGO2 (Hoffart et al., 2011) uses the Reuters newswire articles also used in the CoNLL 2003 shared task on named entity recognition, MEAN-TIME (Minard et al., 2016) uses Wikinews articles, RSS-500 (Röder et al., 2014) contains articles from international

newspapers, OKE2015 (Nuzzolese et al., 2015) uses sentences from Wikipedia articles, the different Entity Discovery and Linking (EDL) tasks from the TAC campaigns use newswires along with broadcast conversation or discussion forums (Ji et al., 2010). Concerning social media, (Rizzo et al., 2015) proposed the NEEL2015 Microposts dataset composed of event-annotated tweets. One can note that all the aforementioned datasets link entities to general-domain knowledge bases, typically DBpedia or Freebase. More similar to our work, Dai et al. (2018) proposed to link mentions to entities defined in a specific social media KB. They constructed the Yelp-EL dataset from the *Yelp* platform where mentions in *Yelp* reviews are linked to *Yelp* business entities, but this corpus does not include visual information.

Entity linking on tweets. While the EL task was initially defined for documents such as newspaper articles, it was also applied to new textual forms such as tweets. In this context, collective approaches such as (Huang et al., 2014) use mention co-referencing to collectively resolve them while Liu et al. (2013) measure *mention-mention* and *mention-entity* similarities from groups of related tweets. Shen et al. (2013) propose a graph-based approach to model the interaction between the topic of interest of Twitter users and all KB entities. Other approaches combine user’s social features (user’s interest and popularity) and temporal reasoning such in (Hua et al., 2015). Fang and Chang (2014) and Chong et al. (2017) extend the context of a target mention to tweets that are close in space and time to the tweet of the target mention.

Multimodal corpora. Many corpora provide images with associated textual content, in particular for the tasks of automatic image annotation (Young et al., 2014; Ginsca et al., 2015), cross-media retrieval (Karpathy and Fei-Fei, 2015; Tran et al., 2016a), image-sentence matching (Hodosh et al., 2013; Ordonez et al., 2011), text illustration (Feng and Lapata, 2010; Chami et al., 2017) and cross-media classification (Tran et al., 2016b; Tamaazousti et al., 2017). Most corpora used in this context consist in images with captions from Flickr (Ordonez et al., 2011; Hodosh et al., 2013; Young et al., 2014) or using Amazon’s Mechanical Turk (Rashtchian et al., 2010; Lin et al., 2014). Other authors nevertheless preferred to build datasets of images with captions in a real context, that is to say extracted from real news articles (Feng and Lapata, 2010; Tirilly et al., 2010; Hollink et al., 2016). In terms of size, these datasets contain from few thousands (Feng and Lapata, 2010; Rashtchian et al., 2010; Hodosh et al., 2013) to several hundred thousand (Hollink et al., 2016) and even one million of captioned images (Ordonez et al., 2011). Other corpora have also been created for more specific usages such as visual question answering (Antol et al., 2015), visual dialogs (Das et al., 2017) or understanding the interactions and relationships between objects in an image (Krishna et al., 2017). Moon et al. (2018) addressed the task of multimedia entity linking and evaluate their approach on a corpus of 12K user-generated image and textual caption pairs from Snapshat, where mentions are linked to the general-domain Freebase KB. However, the evaluation corpus was not released and the authors provide little information on its char-

acteristics and the method used to build it. To the best of our knowledge, our approach is thus the first to allow to build a dataset for MEL evaluation, with full access to image and corresponding text, and usable for reproducible researches.

3. MEL Task Definition

As in standard EL, MEL aims at mapping ambiguous textual mentions to entities in a KB with the distinctive feature that both mentions and entities are characterized by multimodal data. More specifically, we consider that each tweet containing an ambiguous mention has an associated image, and each entity in the KB is represented by a set of tweets accompanied by images. An entity corresponds to a twitter user u (person or organization), with a unique screen name (e.g., @algore), and a timeline containing all the tweets (text+image pairs) posted by this user. A mention m corresponds to an ambiguous textual occurrence of an entity e_x mentioned in a tweet t by an entity e_y , where $e_x \neq e_y$.

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 its 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 entity $e^*(m_i)$ associated with the mention m_i , which is defined as finding the most similar entity to m_i , according to a given similarity:

$$e^*(m_i) = \operatorname{argmax}_{e_j \in KB} \operatorname{sim}(m_i, e_j) \quad (1)$$

Here, we assume that according to a given similarity measure, the multimodal context of a mention m_i is necessarily close to the multimodal context of its corresponding correct entity e^* .

4. Building a Twitter MEL Dataset

The process to build our MEL dataset is mostly automatic and can, therefore, be applied to generate a new dataset at convenience. As depicted in Figure 2, this process comprises two phases: 1) populating the knowledge base, which includes searching new Twitter users (entities) along with their potential ambiguous entities; 2) creating the evaluation corpus by searching tweets that explicitly mention entities from the KB and applying mention replacement to get tweets with ambiguous mentions (see 4.2.2).

4.1. Collecting Data

Data and user metadata were collected between January and April 2019 using the Twitter official API¹. The Twitter API returns up to 3,200 of the most recent tweets of a user². It also gives access to the meta-information of the user such

as his description, the location of his account, his number of followers, etc. We observed that 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 posts (see Section 1). Consequently, we have drawn inspiration from this usage to elaborate a simple process for both candidate entity and ambiguous mention generations.

4.2. Building the Evaluation Dataset

4.2.1. Candidate Entity Generation

Applying the disambiguation of each mention in a dataset over all the entities in the KB is generally intractable. State-of-the-art EL systems use three main techniques for candidate generation, namely surface form expansion, name dictionary and search engine based methods (Shen et al., 2015). In this work, we propose a simple procedure to jointly generate ambiguous candidate entities and populate the KB. Our basic assumption is to consider users sharing the same last name as potential candidate entities for a given mention. In this way, we ensure to sufficiently enrich the KB with ambiguous entities, thus make the EL task challenging. As a first step, we established a non-exhaustive initial list of Twitter user’s screen names using Twitter lists. A Twitter list is a curated set of Twitter accounts³ 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. In practice, 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 users that share the same last name⁴. Additionally, to accelerate knowledge base expansion, we fed the user search engine with a seed list of most common English last names. In this way, it is more likely to find and add as many users as possible in the KB. Users that have been inactive for a long period, non-English users and non-verified user accounts are filtered out.

Increasing the diversity of entity types. The proposed method described above works well for person names. To ensure more diversity in the entities we collect, we also manually collected entities about organization accounts. Given that acronyms are generally ambiguous, we relied on Wikipedia Acronym Disambiguation Pages (WADP) to form groups of ambiguous (organization) entities that share the same acronym. Although we manually collected them, this procedure can be integrated into the whole building process by automatically verifying that organizations from WADP have a corresponding Twitter account.

4.2.2. Generation of Ambiguous Mentions

After building the KB, we used the collected entities to search for tweets that mention them. The Twitter Search

¹<https://dev.twitter.com>

²https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html

³<https://help.twitter.com/en/using-twitter/twitter-lists>

⁴Only the first 1,000 matching results are available with the Twitter API.

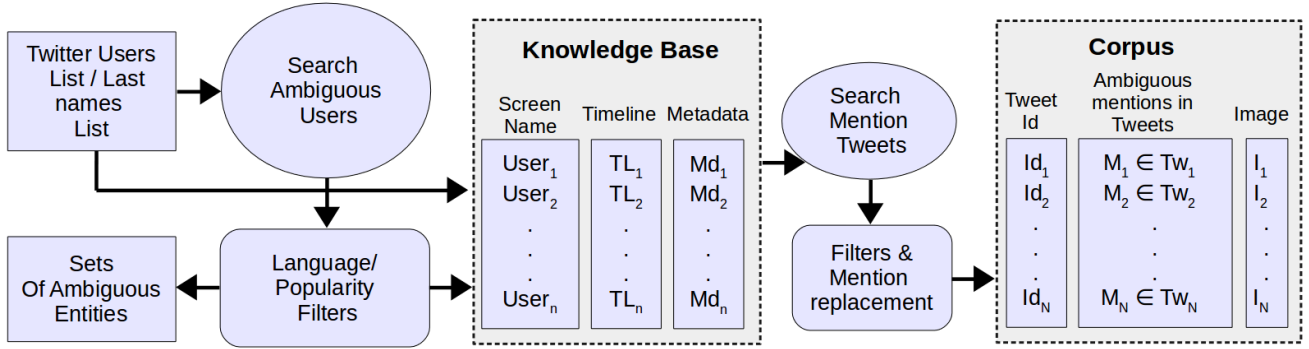


Figure 2: Data collection and corpus construction process.

	# Collected tweets	# Filtered tweets	# Entities or mentions
KB timeline	10M	2.6M	20k
Evaluation corpus	4M	85k	1,678

Table 1: Dataset statistics

API⁵ 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 collected all the retrieved tweets, and 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, in other words, 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 text and image.

5. 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 20.7k entities in the KB, 300 of which are organization entities. Filtering these tweets drastically reduced the size of our data set. Regarding mention tweets, a key factor in the reduction is the elimination of irrelevant tweets where mentions have no linguistic link with the tweet text, i.e., where the mention is included in an enumeration of screen names. Here is a typical example of such tweets: "Hello guys! @user1 @user2 *our-mention* @user3". We considered such cases not comparable to real-world mentions. Discarding these irrelevant tweets as well as tweets without image left a dataset of 2.6M timeline tweets and 20k KB entities. Regarding the evaluation corpus, it left 85k mention tweets covering 1,678 mentions that correspond to 5,571 true grounding entities.

⁵Twitter Search API searches against a sampling of recent tweets published in the past 7 days.

Metrics	Mean	Median	Max	Min	StdDev
# tweets per timeline (text+image)	127.9	52	3,117	1	222.2
# ambiguous entities per mention	16.5	16	67	2	12

Table 2: Statistics on timeline and entity distributions in the MEL dataset

Table 2 shows the timeline tweet distribution of all entities in our KB. As noted by Hua et al. (2015), 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 influences the number of mention tweets we can collect. We necessarily gathered more mention tweets from content generator entities, since they are more likely to be mentioned by others than information seeker entities.

6. Experiments and Results

Having described the proposed approach for building MEL datasets in the context of Twitter and provided some features of our constructed dataset, we empirically evaluate this dataset, and hence our MEL dataset building approach, by reporting the performance of the MEL task on the constructed dataset using an Extra-Trees (Geurts et al., 2006) classifier-based ranking approach. In summary, given a mention m_i and an entity e_i , we respectively extract a fixed continuous representation of their textual contexts using an unsupervised language model. Moreover, the visual contexts of m_i and e_i are determined using a pre-trained convolution neural network employed as an image feature extractor. We additionally provide other traditional features based on popularity and BM25 similarity. We use the Extra-Trees classifier to combine these features, perform classification over the mention-entity pairs and select the best entity among candidate entities.

Textual Context Features

For mention and entity textual context representations, we used the unsupervised Sent2Vec (Pagliardini et al., 2018)



Figure 3: KB entities examples. Each entity in the knowledge base represents a twitter user characterized by its timeline (set of pair of text-image). We assume that combining visual and textual contexts of each entity helps discriminating entities for better EL performance.

sentence embedding model, and more precisely, a pre-trained version on a large Twitter corpus⁶. We adopted this model because we observed, in preliminary experiments, that representations built from the same type of data as ours (social media posts in our case) give better results. 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 second sentence embedding $B_m^{(i)}$ inferred using tweet unigrams and bigrams. Combining their vectors is generally beneficial (Pagliardini et al., 2018). An entity context being represented by a set of tweets, 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.

Visual Context Features

The visual features are extracted with the Inception_v3 model (Szegedy et al., 2016), pre-learned on the 1.2M images of the ILSVRC challenge (Russakovsky et al., 2015). We use its last layer ($D = 1,000$), which encodes high-level information that may help to discriminate between entities. Given an entity e_i , we compute feature vector that is the average of the feature vectors of all the images within its timeline, similarly to the process adopted for the textual context. The visual feature vector of a mention is extracted

from the image of the tweet that contains the mention.

BM25 Features

Given a mention m_i , we compose a bag-of-words vector from all the surrounding words of each occurrence of m_i in mention tweets. In the same way, we compose the bag-of-words vector for a given entity e_i from the words within the entity's timeline tweets. Thereafter, we calculate the cosine similarity of these two vectors weighted by BM25 resulting in a similarity feature.

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 .

6.1. Experimental Setting

We compare in this experiment the results with different Extra-trees (ET) combinations of the following single features:

- Popularity (Pop): baseline feature where the most popular entity is selected, according to the popularity features we considered (the ET classifier is used to combine these features);
- BM25: standard textual context similarity with BM25 weighting;
- S2V-uni: similarity measured between the unigram embeddings extracted using the Sent2Vec sentence model;

⁶<https://github.com/epfml/sent2vec#downloading-pre-trained-models>

Table 3: Multimodal Entity Linking results (mean accuracy and standard deviation over 5 folds) for several approaches. The best value for the hyperparameter p of the Extra-Trees is determined on the training set (with a 4-fold CV) for each of the 5 folds. We report the different values obtained on the second column (and their frequency). For the single-value features, the Extra-Trees is not used. The *popularity* feature vector represents three features thus the performance is estimated with an Extra-Trees

	hyperparameter	Test
Single features		
BM25	-	0.437 / std = 0.0418
S2V-uni	-	0.514 / std = 0.0347
S2V-bi	-	0.521 / std = 0.0306
Img	-	0.294 / std = 0.0338
Combination of features with an Extra-Trees Classifier		
ET(popularity)	p=15 (3/5) p=20 (2/5)	0.568 / std = 0.0892
ET(S2V)	p = 1 (5/5)	0.529 / std = 0.0251
ET(S2V + Img)	p=1 (5/5)	0.541 / std = 0.0270
ET(S2V + Pop)	p=15 (5/5)	0.687 / std = 0.0291
ET(S2V + BM25)	p=1 (5/5)	0.530 / std = 0.0492
ET(S2V + Img + Pop)	p=20 (3/5) p=15 (2/5)	0.691 / std = 0.0381
ET(S2V + Pop + BM25)	p=20 (3/5) p=15 (2/5)	0.718 / std = 0.0474
ET(S2V + Img + Pop + BM25)	p=20 (5/5)	0.723 / std = 0.0391

- S2V-bi: similarity measured between the bigram embeddings extracted using the Sent2Vec sentence model;
- S2V: notation to represent a 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.

To reduce the bias of the classifier performance resulting from a random train/dev/test split validation approach, we adopted a nested k -fold cross-validation procedure to jointly select the best hyper-parameter setting and estimate the average accuracy performance on the whole dataset. In nested k -fold cross-validation, we perform a regular k -fold cross-validation in two nested loops: an inner loop for hyper-parameters tuning and an outer loop for evaluating accuracy performance. More precisely, we split our dataset in k folds, we apply a k' -fold cross validation on the $k - 1$ remaining folds and select the hyper-parameters setting which performed with the best average accuracy score over the k' -folds. Then, we estimate the k^{th} test fold accuracy score. We repeat these steps on the k folds and estimate the global performance by averaging the scores of the k test folds.

Actually, the train/dev/test split does not result from a strict uniformly random draw, since this would bias the dataset toward an easier setting. Indeed, with a uniform draw, most entities would be represented in both the training and validation/testing folds. We considered that to be more realistic, it is important to have some “unique entities” in the validation/testing folds that are not present in the training

one. Beyond realism, this makes the task obviously harder. Hence, we group tweets which refer to the same entity before splitting the dataset into k -folds. This prevents to have all mentions of a test fold to be seen in the training folds. We, therefore, have on average, 50% of unique truth grounding entities in each k^{th} test fold and 80% of unique truth grounding entities in each k^{th} train/dev fold.

The hyper-parameter search space of the Extra-trees classifier with regard to the tree depth is [1, 5, 10, 15, 20, 25]. The hyper-parameters: k is set to 5, k' to 4 and the number of estimators is set to 1000.

6.2. Results

We report in Table 3 the accuracy performance of our classifier for different feature combinations. First, we can see that taking the most popular entity achieves a 56% accuracy score. Recall that our popularity baseline is a combination of a popularity measure represented by the number of followers + friends and a user’s activity measure represented by the number of posted tweets. Thus, the obtained popularity score confirms the hypothesis that popular Twitter users and content generator users are more likely to be mentioned by others. Combining unigram and bigrams similarities ET(S2V) slightly underperforms compared with the popularity baseline. However, integrating all the features in the classifier provides significant performance gains, up to more than 72%. Furthermore, when image similarity is provided (see combinations: ET(S2V + Img), ET(S2V + Img + Pop) and ET(S2V + Img + Pop + BM25)), the performance always increases, demonstrating the interest of considering the visual context in the EL task. We also note that adding image information results in varying performance gains depending on the combined features. This shows the limit of an approach where similarity features are combined using a standard classifier. Indeed, examining the common cor-

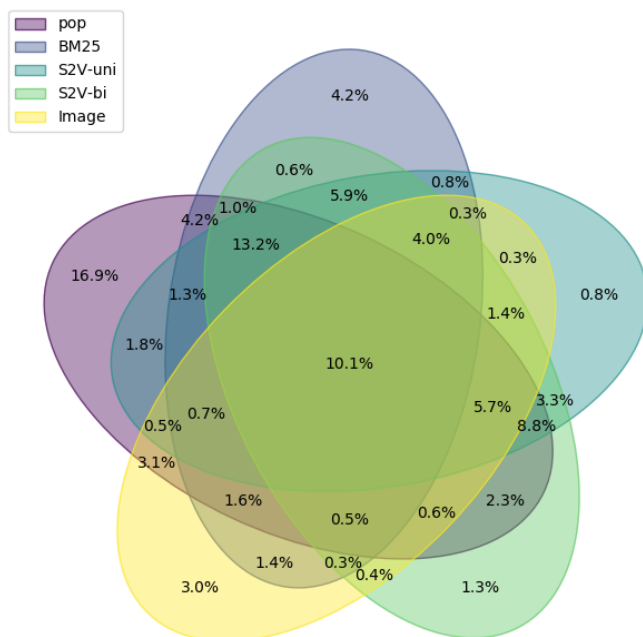


Figure 4: Ratio of common linked entities between single features.

rect linked entities between each feature individually (see Figure 4) shows that the image feature allows linking a significant number of mentions that are not linked by other textual/engineered features. Thus, we believe that integrating textual and visual context representations in a more complex model would better show the influence of image information. Finally, we can reasonably state that our approach for representing the textual/visual mention/entity contexts is relevant, but we plan to investigate a more sophisticated entity representation than the average sentence embedding representation on all the timeline tweets.

7. Conclusion

We propose, in this paper, an approach for building large-scale annotated datasets for the Multimodal Entity Linking task applied to tweets. It entails jointly building a corpus of tweets with ambiguous mentions and a knowledge base with candidate entities. MEL on tweets involves textual ambiguous mentions that are linked to entities representing Twitter users. Both mentions and entities are characterized by textual and visual contexts. The corpus resulting from our proposed method is artificial since we replace Twitter anchors (screen names) with users last names for persons or acronyms for organizations but our approach for creating ambiguous mentions is based on observations of real-world ambiguities which are inherent to the Twitter usage and the generated tweets with ambiguous mentions are similar to real tweets. Preliminary experiments on a dataset built using our approach show its interest. In particular, it offers the opportunity to investigate the benefit of exploiting related visual information to improve the Entity linking task.

8. Acknowledgements

This research was partially supported by Labex DigiCosme (project ANR11LABEX0045DIGICOSME) operated by

ANR as part of the program “Investissement d’Avenir” Idex ParisSaclay (ANR11IDEX000302).

9. Bibliographical References

- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *ICCV*.
- Bunescu, R. and Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*.
- Chami, I., Tamaazousti, Y., and Le Borgne, H. (2017). Amecon: Abstract meta-concept features for text-illustration. In *ICMR*, Bucharest.
- Chong, W.-H., Lim, E.-P., and Cohen, W. (2017). Collective entity linking in tweets over space and time. In *ECIR*. Springer.
- Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL*.
- Dai, H., Song, Y., Qiu, L., and Liu, R. (2018). Entity Linking within a Social Media Platform: A Case Study on Yelp. In *EMNLP*.
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual dialog. In *CVPR*.
- Dredze, M., McNamee, P., Rao, D., Gerber, A., and Finin, T. (2010). Entity disambiguation for knowledge base population. In *COLING*.
- Fang, Y. and Chang, M.-W. (2014). Entity linking on microblogs with spatial and temporal signals. *Transactions of the Association for Computational Linguistics*, 2.
- Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *NAACL-HLT, HLT’10*, Los Angeles, California. Association for Computational Linguistics.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1).
- Ginsca, A. L., Popescu, A., Le Borgne, H., Ballas, N., Vo, P., and Kanellos, I. (2015). Large-scale image mining with flickr groups. In *International Conference on Multimedia Modelling (MMM)*.
- Globerson, A., Lazic, N., Chakrabarti, S., Subramanya, A., Ringaard, M., and Pereira, F. (2016). Collective entity resolution with multi-focal attention. In *ACL*.
- Guo, Z. and Barbosa, D. (2014). Entity linking with a unified semantic representation. In *WWW*. ACM.
- Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*.
- Hoffart, J., Yosef, M. A., Bordino, I., Fürstena, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *EMNLP*. Association for Computational Linguistics.
- Hollink, L., Bedjeti, A., van Harmelen, M., and Elliott, D. (2016). A corpus of images and text in online news. In *LREC*, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Hua, W., Zheng, K., and Zhou, X. (2015). Microblog entity linking with social temporal context. In *ACM SIGMOD*. ACM.

- Huang, H., Cao, Y., Huang, X., Ji, H., and Lin, C.-Y. (2014). Collective tweet wikification based on semi-supervised graph regularization. In *ACL*.
- Ji, H., Grishman, R., Dang, H. T., Griffitt, K., and Ellis, J. (2010). Overview of the TAC 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*.
- Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International Journal of Computer Vision*, 123(1), May.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P. (2014). Microsoft coco: Common objects in context. In *ECCV*.
- Ling, X., Singh, S., and Weld, D. S. (2015). Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3.
- Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., and Lu, Y. (2013). Entity linking for tweets. In *ACL*, volume 1.
- Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *CIKM*. ACM.
- Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., Van Erp, M., Schoen, A., and Van Son, C. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In *LREC*.
- Moon, S., Neves, L., and Carvalho, V. (2018). Multimodal Named Entity Disambiguation for Noisy Social Media Posts. In *ACL*.
- Nuzzolese, A. G., Gentile, A. L., Presutti, V., Gangemi, A., Garigliotti, D., and Navigli, R. (2015). Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*. Springer.
- Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, et al., editors, *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc.
- Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL-HLT*.
- Pershina, M., He, Y., and Grishman, R. (2015). Personalized page rank for named entity disambiguation. In *NAACL-HLT*.
- Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanicalturk. In *NAACL-HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*.
- Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *ACL*. Association for Computational Linguistics.
- Rizzo, G., Basave, A. E. C., Pereira, B., Varga, A., Rowe, M., Stankovic, M., and Dadzie, A. (2015). Making Sense of Microposts (# Microposts2015) Named Entity Recognition and Linking (NEEL) Challenge. In *WWW 2015, 5th International Workshop on Making Sense of Microposts (#Microposts'15)*.
- Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (2014). N³-A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *LREC*.
- Rosales-Méndez, H. (2019). Towards better entity linking evaluation. In *2019 World Wide Web Conference (WWW 2019)*, pages 50–55. ACM.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3).
- Shen, W., Wang, J., Luo, P., and Wang, M. (2013). Linking named entities in tweets with knowledge base via user interest modeling. In *ACM SIGKDD*. ACM.
- Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2).
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.
- Tamaazousti, Y., Le Borgne, H., Popescu, A., Gadeski, E., Ginsca, A., and Hudelot, C. (2017). Vision-language integration using constrained local semantic features. *Computer Vision and Image Understanding*, 163(Supplement C). Language in Vision.
- Tirilly, P., Claveau, V., and Gros, P. (2010). News image annotation on a large parallel text-image corpus. In *LREC*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Tran, T. Q. N., Le Borgne, H., and Crucianu, M. (2016a). Aggregating image and text quantized correlated components. In *CVPR*, Las Vegas, USA, june.
- Tran, T. Q. N., Le Borgne, H., and Crucianu, M. (2016b). Cross-modal classification by completing unimodal representations. In *ACM Multimedia 2016 Workshop: Vision and Language Integration Meets Multimedia Fusion*, Amsterdam, The Netherlands, october.
- Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al. (2015). GERBIL: general entity annotator benchmarking framework. In *WWW*.
- Van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., and Waitelonis, J. (2016). Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC*.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *ACL*.