Jointly Linking Visual and Textual Entity Mentions with Background Knowledge

Shahi Dost[1,2](✉), Luciano Serafini[1], Marco Rospocher[3], Lamberto Ballan[2], and Alessandro Sperduti[2]

[1] Fondazione Bruno Kessler, Trento, Italy
sdost@fbk.eu
[2] University of Padova, Padova, Italy
[3] University of Verona, Verona, Italy

Abstract. "A picture is worth a thousand words", the adage reads. However, pictures cannot replace words in terms of their ability to efficiently convey clear (mostly) unambiguous and concise knowledge. Images and text, indeed, reveal different and complementary information that, if combined, result in more information than the sum of that contained in the single media. The combination of visual and textual information can be obtained through linking the entities mentioned in the text with those shown in the pictures. To further integrate this with agent background knowledge, an additional step is necessary. That is, either finding the entities in the agent knowledge base that correspond to those mentioned in the text or shown in the picture or, extending the knowledge base with the newly discovered entities. We call this complex task Visual-Textual-Knowledge Entity Linking (VTKEL). In this paper, after providing a precise definition of the VTKEL task, we present a dataset composed of about 30K commented pictures, annotated with visual and textual entities, and linked to the YAGO ontology. Successively, we develop a purely unsupervised algorithm for the solution of the VTKEL tasks. The evaluation on the VTKEL dataset shows promising results.

Keywords: AI · NLP · Computer vision · Knowledge representation · Semantic web · Entity recognition and linking

1 Introduction

Given the prominent presence in the web of documents that combines text and images, it becomes crucial to be able to properly process them. In spite of the maturity and reliability of natural language processing (NLP) and computer vision (CV) technologies, an independent processing of the textual and visual part of a document is not sufficient. A more integrated process is necessary. Indeed, the pictorial and textual parts of a document typically provide complementary information about a set of entities occurring both in the picture and

© Springer Nature Switzerland AG 2020
E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 264–276, 2020.
https://doi.org/10.1007/978-3-030-51310-8_24

in the text. For instance, in a news about a car accident, the text may mention the brand and model of the car and the name of the driver, while the picture may reveal the car brand and model as well, but also the car color and its status after the crash. The information conveyed by the two media can be joined by linking the entities mentioned in the text with those shown in the pictures, possibly integrating them with some background knowledge that provides further information about the entities. We call this task *Visual-Textual-Knowledge Entity Linking (VTKEL)*. More precisely, the VTKEL task aims at detecting and linking the maximum visual and textual portions of a document that refer to the same or individual entities of the document, a.k.a. *entity mentions*, with the corresponding entity (or a newly created one) in a knowledge base.

State-of-the-art only provides partial solutions to the VTKEL task. Namely entity linking [1] align textual mentions to entities of a knowledge base, coreference resolution [2] links different textual mentions of the same entity, visual entity linking [3] align visual entity mentions to a knowledge base, visual semantic alignment [4] links different visual entity mentions that refer to the same entity, and, text to image coreference [5] aligns visual and textual mentions of the same entity.

The paper introduces VT-LinKEr[1] (Visual-Textual-Knowledge Entity Linker), an algorithm for solving the VTKEL task that combines state-of-the-art NLP and computer vision tools, and ontological reasoning. Given a document composed of text and image, VT-LinKEr applies an object detector to the image, resulting in a set of bounding boxes labeled with classes of the ontology. Each bounding box is called *visual mention* and the corresponding object, which is an instance of the class label, is called *visual entity*. In parallel, VT-LinKEr processes the text with a tool for entity recognition, which labels the noun phrases with classes of the ontology. The recognized noun phrases are called *textual mentions* and the corresponding instances of the ontological class are *textual entities*. Finally, VT-LinKEr attempts to link visual and textual mentions which correspond to the same entity. This final task is done by exploiting ontological knowledge about class/subclass hierarchy, and similarity information available in the textual mentions.

To evaluate VT-LinKEr, we created a ground-truth dataset for the VTKEL task, called the *Visual-Textual-Knowledge Entity Linking dataset (VTKEL)*. This dataset is derived from the Flickr30k-Entities [6] dataset, which contains about 30 K images, each described by 5 captions. Each picture is annotated with bounding boxes for objects and with coreference chains (a coreference chain links mentions of the same entities across different captions with the corresponding bounding box). We extended the Flickr30k-Entities by annotating each element of the coreference chains with the proper ontological class. As a reference ontology, we adopted YAGO [7]. Since the linking of the ontological class is performed automatically (by using PIKES [8,22]), we manually evaluate the accuracy by checking 1000 randomly selected entries from the VTKEL dataset. The resulting accuracy was about 95% (notice that PIKES annotates *all* the noun phrases).

[1] https://github.com/shahidost/Baseline4VTKEL.

Out of the 1000 pictures, we created a dataset, called VTKEL*, by manually correcting the errors.

We evaluate VT-LinKEr on both VTKEL and VTKEL* datasets. The evaluation is performed in three sub-tasks i.e. visual entities detection and typing, textual entities detection and typing, and visual textual coreference. The F1 measure for visual entities detection and typing on VTKEL* and VTKEL is 65.7% and 64.9% respectively; The F1 measure for textual entities detection and typing 91.8% and 90.5%; the F1 for visual textual conference is 57.1% and 50.4%.

The paper is structured as follows: in Sect. 2, we give a detailed formulation of the VTKEL task. In Sect. 3, we review the main approaches related to the VTKEL task and argue that only partial solutions are available. Section 4 describes VT-LinKEr in details. In Sect. 5, we describe the experiments. Section 6 provides some conclusions and future research directions.

2 Visual-Textual-Knowledge Entity Linking

The Visual-Textual-Knowledge Entity Linking (VTKEL) task takes in input a document composed of text and a picture.[2] More precisely, a document d is a pair $\langle d_t, d_i \rangle$, where d_t is a text in natural language represented as a string of characters and d_i is an image, represented as a 3-channel $(w \times h)$-matrix. We ignore all the structural information about the document, e.g. the relative position of the image w.r.t. the text, the explicit references to the figures, etc. If e is an entity of the domain of discourse of a document d, for example a specific *car* or a *person*, a *textual mention* of e in d is a portion of the text d_t that refers to the entity e. Such a mention can be identified by an interval $\langle l, r \rangle$ with $0 \leq l < r \leq len(d_t)$, corresponding to the characters (in d_t) of the mention. Analogously, a *visual mention* of an entity e is a region of the picture d_i that shows (a characterizing part of) the entity e. E.g., the region of a picture that shows the (face of a) *person* is a visual mention of that *person.* If we restrict to rectangular regions (a.k.a. bounding boxes) a visual mention can be represented by a bounding box encoded by four integers $\langle x, y, x+w, y+h \rangle$ with $0 \leq x, x+w \leq width(d_i)$ and $0 \leq y, y+h \leq height(d_i)$, where $\langle x, y \rangle$ represents the position of the pixel in the top left corner of the bounding box, and w, h represent the width and height of the bounding box (in pixels).

A knowledge base is a logical theory that states properties and relations about a set of entities, called the domain, using a logical language. In description logics a knowledge base is composed of a T-box and an A-box. The T-box contains a set of axioms of the form $\mathsf{C} \sqsubseteq \mathsf{D}$ and $\mathsf{R} \sqsubseteq \mathsf{S}$, for some concept expressions C and D and relations R and S stating that C is a subclass of D (R is a sub-relation of S). The A-box contains assertions of the form $C(e)$ (the entity e is of type C) and $R(e, f)$ (the pair of entities $\langle e, f \rangle$ are in relation R) where e and f are

[2] For the sake of simplicity, we consider only documents that contain one single picture. The extension to multiple pictures, though intuitive, presents additional challenges that are out of the scope of this paper.

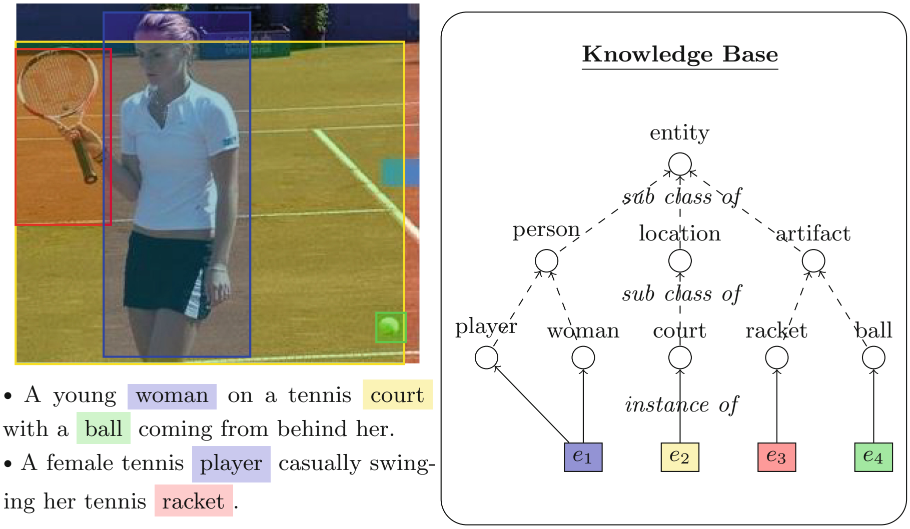
• A young woman on a tennis court with a ball coming from behind her.
• A female tennis player casually swinging her tennis racket.
Knowledge Base
entity
sub class of
person
location
artifact
sub class of
player
woman
court
racket
ball
instance of
e_1
e_2
e_3
e_4

Fig. 1. The picture shows the output of the VTKEL task, which takes in input a picture with related text and an ontology. The output consists of a set of visual mentions (c.f. the bounding boxes in the image) and textual mentions (c.f. the highlighted words in the sentences), corresponding to the mentioned entities (in this case: a ball, a woman, the tennis court and a racket), and the extension (or alignment) of the ontology with entities of the correct (most specific) type.

entities of the knowledge base and C and R are concept and role expressions respectively. The *entities* of a knowledge base are constant symbols that explicitly occur in some axiom of the T-box or assertion of the A-box. For instance, the T-box may contain the knowledge that every car has a manufacturer and that a manufacturer is a company. This knowledge can be formalized by the axioms $\mathsf{Car} \sqsubseteq \exists\, \mathsf{hasManufacturer}.\ \mathsf{Manufacturer}$ and $\mathsf{Manufacturer} \sqsubseteq \mathsf{Company}$, where Car, $\mathsf{Manufacturer}$, and $\mathsf{Company}$ are concept names and $\mathsf{hasManufacturer}$ is a relation (or role). The A-box may contain the knowledge that a specific car (an entity), say car_{22}, is a BMW and that BMW is a $\mathsf{Manufacturer}$. This is formalized by the assertional axioms $\mathsf{Car}(\mathsf{car}_{22})$, $\mathsf{hasManufacturer}(\mathsf{car}_{22}, \mathsf{BMW})$, and $\mathsf{Manufacturer}(\mathsf{BMW})$.

Problem 1 (VTKEL). Given a document composed of a text d_t and an image d_i and a knowledge base K, *VTKEL* is the problem of detecting all the entities mentioned in d_t *and* shown in d_i, and linking them to the corresponding entities in K, if they are present, or to newly added entities of the correct type.

An example of the result of the VTKEL task is shown in Fig. 1. VTKEL is a complex task that requires the solution of a set of well studied elementary tasks in NLP, CV, and logical reasoning. In particular, the following are the key subtasks of VTKEL: entity recognition and classification (i.e. typing) in texts [9]; object detection in images [10]; textual co-reference resolution [2]; textual entity linking to a knowledge base (ontology) [1]; visual entity linking to a knowledge base (ontology) [11]; visual and textual co-reference resolution [4,5,12]. We propose a method to solve the VTKEL task which is obtained by composing state-of-the-art tools that solve some of the subtasks listed above.

3 Related Work

Recently the NLP and CV scientific communities devoted some effort in investigating the interaction and integration of text and image. For an exhaustive survey of the approaches in the area of entity information extraction and linking, we refer the reader to [13]. In particular: [5] exploits natural language descriptions of a picture in order to understand the content of the scene itself. The proposed approach solves the image-to-text coreference problem. It successively exploits the visual information and visual-textual coreference previously found to solve coreference in text. The work described in [14,15] tackles the problem of ranking the concepts from the knowledge base that best represents the core message expressed in an image. This work involves the three elements: Image, Text, and Knowledge, but it does not provide information about the entities mentioned in the text and shown in the image. The approach in [3] adapts Markov Random Fields to represent the dependencies between what is shown in the frames of videos about the wild-life animal and the subtitles. The main objective is to detect the animal shown in a frame, and the mentions of animal in the subtitle. The set of entities are the animal names available in WordNet [16]. Object detection is not performed: the approach assumes that only one animal is shown in a frame, and the vision part consists of image classification. Furthermore, no background knowledge about animals is used. [11] proposes a basic framework for visual entity linking to DBpedia and Freebase. The approach involves also textual processing since the link of bounding boxes to DBpedia and Freebase entities is found passing through an automatically generated textual description of the image. The approach uses the Flickr8k dataset, which is a subset of the Flikr30k-Entities dataset. A combination of textual coreference resolution and linking of image and textual mentions is described in [17] with the objective of solving the problem of assigning names to people appearing in TV-show.

Concerning datasets that combine text and images, there are several resources available, however, none of them have all the three components necessary for the VTKEL task. VisualGenome [18] is an extremely large dataset that contains pictures in which objects are annotated with their types, attributes, and relationships. Annotations are mapped to WordNet synsets. Objects can also be annotated with some short sentence that describes some qualitative property of the object. E.g., “The girl is feeding the elephant” or “a handle of bananas”. However, there is no alignment between the objects mentioned in these phrases and the objects shown in the picture. E.g., there is no bounding box for the object “bananas” or “elephant”. The Visual Relationship Dataset (VRD) [19] is a dataset of images annotated with bounding boxes around key objects. Furthermore, VRD contains annotations about relationships between objects in the form of triplets $\langle \text{object\_type}, \text{relation}, \text{subject\_type} \rangle$ describing the scene. Examples of annotations are $\langle man, riding, bicycle \rangle$ and $\langle car, on, road \rangle$. However, these annotations are not aligned to any knowledge base. The Microsoft COCO dataset [20] contains pictures associated with five captions. They are annotated with objects regions of any shape (not simple bounding boxes) and each region is assigned with an object-type. This dataset does not contain any information about the relation between object regions, and the relation between regions and mentions in the captions. Conceptual Captions [21] is a recently introduced dataset that has been developed for automatic image caption generation. It contains one order of magnitude more items than Microsoft COCO. It is a realistic dataset as images with captions have been automatically extracted and filtered from the web. However, there is no visual/textual mention annotation and visual textual entity linking.

From the above analysis, it becomes clear that there is not a single, comprehensive approach corresponding to the VTKEL task. This justifies the introduction of the task, the development of a ground truth dataset, and a first (baseline) algorithm for its solution.

4 The VT-LinKEr Algorithm

VT-LinKEr is composed of two sequential phases: The first phase, *the entity detection phase*, focuses on visual entity detection & typing (VMD-VET) and textual entity detection & typing (TMD-TET); the second phase, *the matching phase*, attempts to match the discovered entities i.e. visual-textual coreference (VTC). The entity detection phase is based on the output of state-of-the-art tools in NLP and CV. The matching phase is realized by using the semantic matching which exploits the knowledge available in the T-box (i.e. class/subclass hierarchy). In the following, we illustrate the different steps for each phase.

Visual Mention Detection (VMD): To implement VMD, we process images with YOLO [23], which returns a set of bounding box proposals each of which is associated with a YOLO-class and a confidence score in [0,1]. We used the model pre-trained on the 80 classes of the COCO dataset. Among the bounding box candidates, we retain only those having confidence equal or greater than a specified threshold (in the experiments we set it to 0.5). In general, one could use some more sophisticated selection criteria that take into account also the co-occurrence with the other bounding box candidates (e.g., glass and bottle are more probable than glass and elephant) and the output of the textual mention detection and the ontological knowledge. For the picture of Fig. 1, YOLO returns three bounding box candidates with score higher than 0.5, labeled with *person*, *ball* and *racket*, but no bounding box has been found for the tennis court (due to the lack of appropriate classes for locations in the YOLO class set).

Visual Entity Typing (VET): The objective of this sub-task is to find the correct most specific class in the knowledge base that can be associated to each visual entity associated to the visual mention detected in the VMD step. Notice that the COCO class does not correspond one-to-one with the YAGO classes, this implies that we need to map the class returned by YOLO into YAGO. A naïve way to implement this task is to map the label contained in the output of the object detector to its corresponding ontology class. Also, here more sophisticated methods can be implemented that take into account also the weight of the labels or additional visual/numerical features. In the VT-LinKEr algorithm, we adopt the straightforward approach of manually mapping the 80 COCO classes to the corresponding (most specific) classes of the YAGO ontology.[3] Examples of mappings from COCO to YAGO are: *person* → `yago:Person100007846`, *ball* → `yago:Ball102778669`, and *hotdog* → `yago:Frank107676602`.

Textual Mention Detection (TMD): To detect textual mentions of entities we process the text with the PIKES suite, which provides services for both textual mention detection and textual entity typing to the YAGO ontology. These two tasks are tightly integrated in PIKES, however, for conceptual clarity, here we present them separately. Let us focus on the entity mention detection. Given a text in input PIKES applies different state-of-the-art NLP techniques to discover entity mentions depending on their "nature":

- *named entity mentions* (e.g., Barak Obama, Trento, IBM) refers to entities for which there is an individual in the knowledge base. They are recognized and linked (performing a task called Entity Linking) to the corresponding entity in YAGO (the knowledge base is not extended).
- *common nouns* (e.g., racket, ball, player, and woman) implicitly identify entities, by referring to their type (e.g., the mention of "racket" does not refer to the general notion of racket, but to a specific object, of type racket). Common nouns are discovered via word sense disambiguation (WSD). For every common noun, WSD returns the WordNet synset corresponding to the correct sense in which the noun is used. For instance, the correct sense of "racket" is the one indicating a sport equipment, and not a loud and disturbing noise. A new entity is created and added to the knowledge base for common nouns occurring in the text.

  Some further processing is performed to properly handle compound noun phrases (e.g., "a female tennis player"). PIKES also performs a syntactic analysis of the text: in particular, words in a noun phrase can be tagged either with *head* or with *modifier*, depending on their syntactic role in the noun phrase (e.g., in "a female tennis player" the noun "player" is the head and "female" and "tennis" are modifiers). In the current version of the VT-LinKEr algorithm, a new entity is added to the knowledge base only for the head noun, and not for its modifiers.

For example, for the first sentence of the caption in Fig. 1, PIKES detects three textual mentions: *woman*, *court* and *ball*.

[3] The whole mapping can be downloaded from https://figshare.com/articles/YOLO_to_YAGO_classes_mapping/8889848.

Textual Entity Typing (TET): This task is also implemented using PIKES primitives. Typing for named entities is not necessary since these entities are in the YAGO knowledge base, and thus already typed according to the YAGO ontology. For the common nouns, we exploit the mapping from WordNet to YAGO also available in PIKES to obtain the (more specific) YAGO class associated to the WordNet synset of the mention, and the corresponding type assertion will be added to the knowledge base. For example, for the first sentence of the caption in Fig. 1, PIKES types the entities corresponding to the textual mentions *woman*, *court* and *ball*, with the YAGO classes `yago:Woman110787470`, `yago:Court108329453`, and `yago:Ball102778669`, respectively.

Visual Textual Coreference (VTC): This is the last sub-task that has to be accomplished by VT-LinKEr. For this task, we exploit the class/subclass hierarchy between the classes in the knowledge base. Let VE and TE be the set of textual and visual entities that are mentioned in a visual-textual document, and that are present in the knowledge base with a given type. The coreference sub-task has the objective of finding the coreference relation $CR \subseteq VE \times TE$ such that the following consistent properties hold:

(i) For every $ve \in VE$ there is at least one $\langle ve, te \rangle \in CR$;
(ii) For every $ve \in VE$ there is at most one $\langle ve, te \rangle \in CR$;
(iii) If $\langle ce, ve \rangle$ (ce is the coreference entity) and ve and te are of type C_v and C_t respectively then either $C_v \sqsubseteq C_t$ or $C_t \sqsubseteq C_e$ holds in the knowledge base.

In simple situations, the above criteria uniquely defines the coreference relations. This is the case for instance for the example presented in Fig. 1. However, in many cases the relation $CR \subseteq VE \times TE$ is not uniquely defined by the above criteria. Nevertheless, the problem can be straightforwardly encoded as a MaxSat problem. In case of CRs with equal total weight, a random choice is taken although additional heuristics could be implemented either by using some supervised learning method or by handcrafting the weight of a pair $\langle ve, te \rangle$ by exploiting some additional features of the mentions of ve and te.

5 Experimental Evaluations

To evaluate the performance of VT-LinKEr, we have developed two ground truth datasets [25]. The first one, called VTKEL, has been derived from Flickr30k-Entities, and it is generated automatically by typing the visual and textual entities with classes from the YAGO ontology. The second one, called VTKEL*, has been obtained by randomly selecting 1000 pictures (and the corresponding captions) from the first dataset, and manually validating and revising the proposed alignments to YAGO. In the following, we provide some details on the datasets, and then we describe the evaluations conducted.

5.1 Datasets

The first dataset called VTKEL, [4] has been obtained by extending the Flickr30k-Entities dataset by linking textual and visual mentions to entities assigned with the most specific YAGO class. Looking at Fig. 1, we started form the left part of the figure (the picture and captions, with annotated visual and textual mentions, and alignment between corresponding mentions), available in the Flickr30k-Entities, and we extended it with the right part, by populating a knowledge base with corresponding entities typed according to the YAGO ontology. The 30K VTKEL dataset has been automatically produced by processing the captions of Flickr30k-Entities with PIKES for entity recognition and linking to YAGO. Specifically, for each textual mention (aligned to a visual mention) in Flickr30k-Entities, detected also by PIKES, a corresponding entity is created (or aligned to, if already existing) and typed according to the appropriate YAGO ontology.

The second dataset, called VTKEL*,[5] has been obtained by randomly sampling 1000 entries from the VTKEL dataset (corresponding to $20,356$ textual mentions, and 8673 visual mentions). Every entry of VTKEL* has been manually checked for the correctness and completeness of the YAGO classes associated to the mentioned entities. Wrong and missing links are manually adjusted. Errors are mainly due to the incorrect word sense disambiguation: e.g., in some cases, "bus" was linked to the concept of the computer bus, instead of that of coach, and "arm" to weapon instead of bodypart. The construction of VTKEL*-dataset allows us also to estimate the error rate of the larger VTKEL dataset. In particular, we found no missing link (i.e., recall is 100%) and 916 incorrectly linked mentions, which amounts to $Precision = 0.955, Recall = 0.893, and\ F1 = 0.923$. We believe that an error rate of $\approx 4.5\%$ is physiological also in manually developed datasets, and therefore we believe that the VTKEL-dataset can be reasonably considered a ground truth.

To maximize reusability and connection with the Semantic Web, we represent the datasets in RDF. This representation will also support semantic visual query answering via standard SPARQL language. To organize the dataset, we adopt the model proposed in [8], extending it for representing visual mentions. The model is organized in three distinct yet interlinked representation layers: *Resource*, *Mention*, and *Entity layer*.

5.2 Evaluation

We evaluated the performances of VT-LinKEr on both VTKEL and VTKEL* datasets. We separately assessed the performance on the three sub-tasks described in Sect. 4. We use the standard metrics, namely precision (P), recall (R), and F-score (F_1). The figures obtained from the evaluation are reported in Table 1.

[4] The VTKEL dataset can be downloaded from https://figshare.com/articles/VTKEL_dataset/9816242/4.

[5] https://figshare.com/articles/VTKEL_dataset/10318985.

Table 1. VT-LinKEr evaluation results

	VTKEL* dataset			VTKEL dataset		
Task	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
VMD + VET	0.765	0.574	0.657	0.731	0.585	0.649
TMD + TET	0.954	0.884	0.918	0.942	0.872	0.905
VTC	0.586	0.558	0.571	0.514	0.486	0.504

Visual Entities Detection and Typing (VMD) + (VET): To evaluate the visual detection part, we use standard method adopted for evaluating object detection. A visual mention b_p of type t_p produced by VT-LinKEr on an image is considered to be correct if the ground truth annotation of the image contains a bounding box b_g of type t_g such that the intersection over union ratio $(\frac{area(b_p \cap b_g)}{area(b_p) \cup area(b_g)})$ is greater or equal to $\frac{1}{2}$ and if the predicted type t_p is equal or a subclass of t_g in YAGO. For the 1000 entries dataset VTKEL*, VT-LinKEr predicted 6914 total visual entities with respect to the 9243 annotated visual entity objects. VT-LinKEr correctly predicted 5306 ($P = 0.767, R = 0.574, F1 = 0.657$) of them. By using the same procedure for 30 K entries VTKEL dataset, VT-LinKEr predicted 220, 853 total visual entities with respect to the 275, 770 annotated visual entity objects. VT-LinKEr correctly predicted 161, 342 ($P = 0.731, R = 0.585, F1 = 0.649$) of them. In the majority of the cases, VT-LinKEr framework ignored human bodyparts and clothing during the prediction of visual mentions due to the 80 classes of COCO dataset [20]. In some cases, VT-LinKEr predicts additional correct visual mention not annotated in Flickr30k-Entities. In the evaluation, these are considered errors though they are not strictly so.

Textual Entities Detection and Typing (TMD) + (TET): To evaluate the performance of this sub-task, we apply a criterion analogous to the visual entity detection and typing sub-task. A textual mention w_p of an entity of YAGO class t_p predicted by VT-LinKEr on a caption, is considered to be correct if the ground truth annotation on the caption contains a mention w_g of an entity of type t_g such that w_p is equal or a sub-string of w_g and the type t_p is equal or a sub-type of t_g according to the YAGO class hierarchy. From the 5000 captions of VTKEL*dataset, VT-LinKEr wrongly recognized and linked 935 out of total 20, 374 textual entities, which amount to $P = 0.954$, $Recall = 0.884$, and $F1 = 0.918$. Similarly, for 158, 605 captions of VTKELdataset, VT-LinKEr correctly recognized and linked 576, 769 out of total 612, 281 textual entities. Most of the errors during entity recognition and linking are due to the word sense disambiguation.

Visual Textual Coreference (VTC): We evaluate the capability of VT-LinKEr of aligning visual and textual entities. A coreference pair $\langle ve_p, te_p \rangle$ produced by VT-LinKEr is correct, if the ground truth contains the triple

ve_g `owl:sameAs` te_v such that the visual mentions (bounding boxes) of ve_p and ve_g matches (under the IOU ratio), the textual mention of te_p matches the textual mention of te_g (i.e., te_p is equal or a substring of te_g). Notice that here we are not considering the types of the entities. Type compatibility is indeed guaranteed by the fact that coreference pairs are added only if their types are compatible (i.e., they are either equal or in subclass relation in YAGO). From the 1000 entries VTKEL*dataset, VT-LinKEr correctly aligned 4082 visual entities with 8681 textual entities out of total 6914 visual and 14, 786 textual entities ($P = 0.586, R = 0.558, F1 = 0.571$). Similarly, for VTKEL dataset, VT-LinKEr correctly aligned 118, 502 visual entities with 243, 831 textual entities out of total 220, 853 visual and 576, 769 textual entities. In most of the cases, the alignment of human-body parts and clothing with visual entities are missed by VT-LinKEr.

6 Conclusion and Future Works

In this paper, we have introduced a new complex task for recognizing mentions of entities in multimedia documents composed of image and text, and align them with a reference ontology. This task turns out to be rather important for many applications in the area of multimedia indexing processing and retrieval, e.g., information extraction from multimedia systems [24], for visual question answering [26], and for visual textual dialogue systems [27]. We argue that there are advantages to solve the VTKEL task in a collective manner, i.e., trying to jointly solve all the tasks involved in it. For this reason, we created a new dataset annotated with all the information necessary for the VTKEL task. We perform this in a completely automatic manner, by processing the captions of the Flickr30k dataset to find entities and linking them to the YAGO ontology. We also developed the first algorithm to solve the task of VTKEL. The proposed algorithm is developed by using state-of-the-art tools for object detection in images, entity recognition in text, entity linking to ontologies and alignment of visual-textual entity mentions. This allows us to close the loop between language, vision, and knowledge. In the future, we are planning to improve the accuracy of every single sub-task, especially the object detection, by using a more complete set of object classes. We also planned to implement a more sophisticated method for the visual-textual entity matching, based on supervised methods, or statistical relational learning methods. We also want to apply the method to a dataset that includes more pictures and text different from captions (e.g., short news with pictures).

References

1. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. KDE **2**(27), 443–460 (2015)
2. Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R.: Anaphora and Coreference Resolution: A Review. arXiv preprint arXiv:1805.11824 (2018)

3. Venkitasubramanian, A.N., Tuytelaars, T., Moens, M.-F.: Entity linking across vision and language. Multimed. Tools Appl. 1–24 (2017). https://doi.org/10.1007/s11042-017-4732-8
4. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE-CVPR, pp. 3128–3137 (2015)
5. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. In: Proceedings of the IEEE-CVPR, pp. 3558–3565 (2014)
6. Plummer, B.A., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
7. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of WWW 2007, pp. 697–706, May 2007
8. Corcoglioniti, F., Rospocher, M., Aprosio, A.P.: Frame-based ontology population with PIKES. IEEE Trans. KDE **28**(12), 3261–3275 (2016)
9. Goyal, A., Gupta, V., Kumar, M.: Recent named entity recognition and classification techniques: a systematic review. Comput. Sci. Rev. **29**, 21–43 (2018)
10. Han, J., Zhang, D., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE SPM **35**, 84–100 (2018)
11. Tilak, N., Gandhi, S., Oates, T.: Visual entity linking. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 665–672. IEEE, May 2017
12. Huang, D.A., Fei-Fei, L., Carlos Niebles, J.: Unsupervised visual-linguistic reference resolution in instructional videos. In: IEEE-CVPR, pp. 2183–2192 (2017)
13. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: a survey. Semantic Web (Preprint), pp. 1–81 (2018)
14. Weiland, L., Hulpus, I., Ponzetto, S.P., Dietz, L.: Using object detection, NLP, and knowledge bases to understand the message of images. In: Amsaleg, L., Guðmundsson, G.Þ., Gurrin, C., Jónsson, B.Þ., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10133, pp. 405–418. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51814-5_34
15. Weiland, L., Hulpu, I., Effelsberg, W., Dietz, L.: Knowledge-rich image gist understanding beyond literal meaning. DKE **117**, 114–132 (2018)
16. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
17. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with “their” names using coreference resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 95–110. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_7
18. Krishna, R., Zhu, Y., Kravitz, J., Bernstein, M.S.: Visual genome: connecting language and vision using crowdsourced dense image annotations. IJCV **123**, 32–73 (2017)
19. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
20. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
21. Sharma, P., Ding, N., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL, pp. 2556–2565 (2018)

22. Corcoglioniti, F., Rospocher, M., Mostarda, M., Amadori, M.: Processing billions of RDF triples on a single machine using streaming and sorting. In: ACM-SAC (2015)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: IEEE-CVPR, pp. 779–788 (2016)
24. Bracamonte, T., Schreck, T.: Extracting semantic knowledge from web context for multimedia IR: a taxonomy, survey and challenges. In: MTA, pp. 13853–13889 (2018)
25. Dost, S., Serafini, L., Rospocher, M., Ballan, L., Sperduti, A.: VTKEL: a resource for visual-textual-knowledge entity linking. In: Proceedings of ACM Symposium on Applied Computing, pp. 2021–2028 (2020)
26. Antol, S., et al.: VQA: visual question answering. In: IEEE-ICCV, pp. 2425–2433 (2015)
27. Das, A., et al.: Visual dialog. In: Proceedings of the IEEE CVPR, pp. 326–335 (2017)