ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200442

VT-LINKER: Visual-Textual-Knowledge Entity Linker

Shahi Dost[1], **Luciano Serafini**[2], **Marco Rospocher**[3], **Lamberto Ballan**[4] and **Alessandro Sperduti**[5]

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 by 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, we precisely define the VTKEL task and present two datasets composed of 1k and 30k pictures, annotated with visual and textual entities and linked to the YAGO ontology. Successively, we develop the first unsupervised algorithm for the solution of VTKEL task. The evaluation of the algorithm shows promising results on both 1k and 30k VTKEL datasets.

1 INTRODUCTION

Despite 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 in the text. 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).

Problem. *Given a document composed of a text d_t and an image d_i; given a knowledge base K, the Visual-Textual-Knowledge Entity Linking* VTKEL *problem 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.*

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 [3]; object detection in images [4]; textual co-reference resolution [8]; textual entity linking to a knowledge base (ontology) [7]; visual entity linking to a knowledge base (ontology) [9]; visual and textual co-reference resolution [5]. We propose an unsupervised algorithm called VT-LINKER (Visual-Textual-Knowledge Entity Linker) to solve the VTKEL task.

[1] Fondazione Bruno Kessler & University of Padova, Italy, sdost@fbk.eu
[2] Fondazione Bruno Kessler, Italy, serafini@fbk.eu
[3] University of Verona, Italy, marco.rospocher@univr.it
[4] University of Padova, Italy, lamberto.ballan@unipd.it
[5] University of Padova, Italy, alessandro.sperduti@unipd.it

2 THE VT-LINKER ALGORITHM

The VT-LINKER[6], combines state-of-the-art NLP and CV tools, and ontological reasoning for solving the *VTKEL* task. 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 (vm)* and the corresponding object, which is an instance of the class label, is called *visual entity (ve)*. 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* (tm is a textual mention) and the corresponding instances of the ontological class are *textual entities* (te is a textual entity). 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/sub-class hierarchy, and similarity information available in the textual mentions.

Specifically, given a knowledge base, for every input document, composed of an image and some text explaining the content of the image, VT-LINKER produces a set of assertions (RDF triples to be added to the A-box of the knowledge base), each belonging to one of the following five types:

VMD *Visual mention detection triples:* $\langle e, \texttt{isDenotedBy}, vm\rangle$, the visual entity e is denoted by vm.
VET *Visual entity typing triples:* $\langle e, \texttt{hasType}, c\rangle$, the visual entity e which is correspond to a vm is an instance of the knowledge base concept c.
TMD *Textual mention detection triples:* $\langle e', \texttt{isDenotedBy}, tm\rangle$, the entity e' is denoted by textual mention tm (tm is a portion of the text).
TET *Textual entity typing triples:* $\langle e', \texttt{hasType}, c\rangle$, the entity e', corresponding to a tm is an instance of the knowledge base concept c.
VTC *Visual Textual Coreference triples:* $\langle e, \texttt{sameAs}, e'\rangle$, the two entities e and e' denotes the same real-world entity.

How to produce the above assertions is illustrated in the following:

Visual entities detection and typing (VMD) + (VET): To implement VMD, we process images with YOLO [6], which returns a set of bounding box proposals, each of which is associated with a

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

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 $confidence >= 0.5$.

VET finds the correct specific class in the knowledge base associated with each visual entity, via the visual-mention detected by the VMD step. In the VT-LINKER algorithm, we adopt the approach of manually mapping the 80 COCO classes to the corresponding (most specific) classes of the YAGO[7] ontology.

Textual entities detection and typing (TMD) + (TET): To detect textual mentions of entities, we process the text with the PIKES[8] [1] suite, which provides services for both textual mention detection and textual entity typing to the YAGO ontology. PIKES applies different state-of-the-art NLP techniques to discover entity mentions and to link them to YAGO classes. For the common nouns (e.g., woman), we exploit the mapping from WordNet to YAGO to obtain the (more specific) YAGO class associated with the WordNet synset of the mention.

Visual textual coreference (VTC): For this task, we exploit the class/sub-class 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:

1. For every $ve \in VE$ there is at least one $\langle ve, te \rangle \in CR$;
2. For every $ve \in VE$ there is at most one $\langle ve, te \rangle \in CR$;
3. If $\langle ce, ve \rangle$ (ce is 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.

3 EXPERIMENTAL EVALUATIONS

To evaluate the performance of VT-LINKER, we have developed two ground truth datasets [2]. The first dataset, called VTKEL[9], has been obtained by extending the Flickr30k-Entities dataset by linking textual and visual mentions to entities assigned with the most specific YAGO class. 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 tm (aligned to a vm) in Flickr30k-Entities, also detected 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*[10], 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 "arm" was linked to the concept of *weapon* instead of *bodypart*. The construction of the 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 an accuracy of 95%. We believe that an error rate of 5% is physiological also in manually developed datasets, and therefore we believe that the VTKEL-dataset can be reasonably considered a ground truth.

We evaluated the performances of VT-LINKER on both VTKEL* and VTKEL datasets. We separately assessed the performance on the sub-tasks described in Section 2. 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.

Table 1. VT-LINKER evaluation results

	VTKEL* dataset			VTKEL dataset		
Task	P	R	F_1	P	R	F_1
VMD + VET	0.748	0.574	0.650	0.731	0.585	0.649
TMD + TET	0.955	0.884	0.918	0.942	0.872	0.905
VTC	0.586	0.558	0.571	0.514	0.486	0.504

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 sub-class of t_g in YAGO.

TMD + TET: For this sub-task, we apply a criterion that a textual mention w_p of an entity of YAGO class t_p predicted by VT-LINKER on a caption is 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 with respect to the YAGO class-hierarchy.

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 $\langle ve_g \texttt{ owl:sameAs } te_v \rangle$ 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 w.r.t YAGO.

REFERENCES

[1] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio, 'Frame-Based Ontology Population with PIKES', *IEEE Transactions on Knowledge and Data Engineering*, **28**(12), 3261–3275, (2016).

[2] Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti, 'VTKEL: A resource for Visual-Textual-Knowledge Entity Linking', in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, (2020).

[3] Gupta V. Goyal A. and Kumar M., 'Recent named entity recognition and classification techniques: a systematic review', *SC-Review*, (2018).

[4] Cheng G. Han J., Zhang D., 'Advanced deep-learning techniques for salient and category-specific object detection: a survey', *IEEE SPM18*.

[5] Andrej Karpathy A. and Fei-Fei .L, 'Deep visual-semantic alignments for generating image descriptions', in *IEEE-CVPR15*, pp. 3128–3137.

[6] Girshick R. Redmon J., Divvala S. and Farhadi A., 'You only look once: Unified, real-time object detection', in *IEEE-CVPR16*, pp. 779–788.

[7] Wang J. Shen W. and Han J., 'Entity linking with a knowledge base: Issues, techniques, and solutions', *IEEE KDE*, **27**, 443–460, (2014).

[8] Poria S. Sukthanker R. and Cambria E., 'Anaphora and coreference resolution: A review', *arXiv preprint arXiv:1805.11824*, (2018).

[9] Gandhi S. Tilak N. and Oates T., 'Visual entity linking', in *IJCNN*, pp. 665–672, (2017).

[7] http://yago-knowledge.org/
[8] http://pikes.fbk.eu/
[9] https://figshare.com/articles/VTKL_dataset_file/7882781
[10] https://figshare.com/articles/VTKEL_dataset/10318985