AAAA Baseline for Visual-Textual-Knowledge Entity linking

Anonymous EMNLP submission

Abstract

To understand the content of a document containing both text and pictures, an artificial agent needs to jointly recognize the entities shown in the pictures and mentioned in the text, and to link them to its background knowledge. This is a complex task, that we call *visual-textual-knowledge entity linking (VTKEL)*, that aims at linking visual and textual entity mentions to the corresponding entity (or a newly created one) of the agent knowledge base. Solving the VTKEL task open a wide range of opportunities for improving semantic visual interpretation. For instance, given the effectiveness and robustness of state-of-the-art NLP technologies in entity linking, by automatically linking visual and textual mentions of the same entities with the ontology, we can obtain a huge amount of automatically annotated images with detailed categories. In this paper we propose the VTKEL dataset, consisting of images and corresponding captions, in which the image and textual mentions are both annotated with the corresponding entities typed according to the YAGO ontology. The VTKEL dataset can be used for training and evaluating algorithms for visual-textual-knowledge entity linking.

1 Introduction

Understanding the content of documents composed of images and text is nowadays an important task since many of the unstructured documents available on the Web or in internal repositories are composed of text and images that jointly describe one particular topic. The growing maturity and reliability of natural language processing (NLP) and image processing (IP) technologies set the basis for deploying them in many products and real world applications. However the independent processing of the textual and visual part of a document is not sufficient to fully understand its content. A more integrated process is necessary. Indeed, the pictorial and textual parts of a document, while referring to the same entities, typically provide complementary information about them. For instance, in a news about a car accident, the text may mention the brand and model of the car involved in the accident as well as 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. Redundant information between text and images (c.f., the car brand and model) enables matching the visual and textual *mentions* of the same entity (c.f., the car). Matching mentions, in turns allows joining the complementary information (c.f., name of driver, car color and status after crash) contributed independently by the two media. Furthermore, this information is usually interpreted by human agents also in light of some background knowledge. This background knowledge, typically operationalized in terms of a knowledge base (=T-box + A-box), actually plays a double role: on the one side it is used as input for processing and understanding the content of the document; and, on the other side it is augmented with the additional knowledge resulting from the interpretation of the document, i.e., new facts contained in the document about entities either already present or to be added in the background knowledge base.

In this paper we introduce a complex task called *Visual-Textual-Knowledge Entity Linking (VTKEL)*, which aims at linking the visual and textual portions of a document that refer to the same entity, a.k.a. *entity mentions*, with the corresponding entity (or a newly created one) in a knowledge base. We generalize the notion of entity mention typically adopted in NLP (e.g., (Doddington et al., 2004)) to images. More in details, a *visual entity mention* is a region of an image (e.g., a rectangular bounding box as typically considered in IP) that refers to an entity. VTKEL is a mandatory task to support any other information extraction activity involving the entities mentioned in the document. State of the art approaches only provide partial solutions to the problem: entity linking (Shen et al., 2015) aligns textual mentions to entities of a knowledge base; coreference resolution (Sukthanker et al., 2018) links different textual mentions of the same entity; visual entity linking (Venkitasubramanian et al., 2017) aligns visual entity mentions to a knowledge base; visual semantic alignment (Karpathy and Fei-Fei, 2015) links different visual entity mentions that refer to the same entity; and, text to image coreference (Kong et al., 2014) aligns visual and textual mentions of the same entity. However none of the above approaches tackle the problem as a whole task, aligning textual, visual, and background knowledge content.

The second contribution of the paper is the assembling of a ground-truth dataset for the VTKEL task. We introduce the *Visual-Textual-Knowledge Entity Linking* (VTKEL) dataset. VTKEL, derived from Flickr30k-Entities (Plummer et al., 2015), consists of documents composed of a picture and five corresponding descriptions captioning it. VTKEL was automatically derived by (i) applying a knowledge graph extraction tool (PIKES (Corcoglioniti et al., 2016)) to the textual captions of the Flickr30k dataset, and (ii) leveraging the picture-caption coreference annotations contained in the original Flickr30k dataset. As a result, visual and textual mentions of each picture and captions are annotated and aligned to entities typed with classes from Yago (Suchanek et al., 2007), a well-known Semantic Web (SW) ontology. Such dataset is essential for providing training and evaluation material for automatic algorithms tackling the VTKEL task.

2 Visual-Textual-Knowledge Entity Linking task

The Visual-Textual-Knowledge Entity Linking (VTKEL) task takes in input a document composed of text and a picture.[1] 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. Notice that, for the sake of simplicity, 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 characterising 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 K is a logical theory, expressed for instance in some language of Description Logics, composed of a T-box and an A-box. The T-box contains a set of axioms of the form $CisaD$ and $RisaS$, for some concept expression C and D and role expressions R and 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 entities of the knowledge base and C and R are concept and role expressions respectively. A Knowledge Base K expresses, in a logical form, the knowledge about a domain which is populated by a set of entities. The *named entities* of a knowledge base K are the individual constants that appears in some axioms of the T-box of K or in the assertions of the A-box of K. A knowledge base can contain also entities which are not explicitly named. For instance if K contains an axioms of the form $AisaexistsR.B$ (intuitively: every entity of type A is related via R with an entity of type B), and the A-box of K contains the fact $A(e)$, then there should exist an entity (possibly not named) that is in relation R with e and that is of type B.

Problem 1 (Visual-Textual-Knowledge Entity Linking) *Given a document d composed of a text d_t and an image d_i and a knowledge base K,* Visual-

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

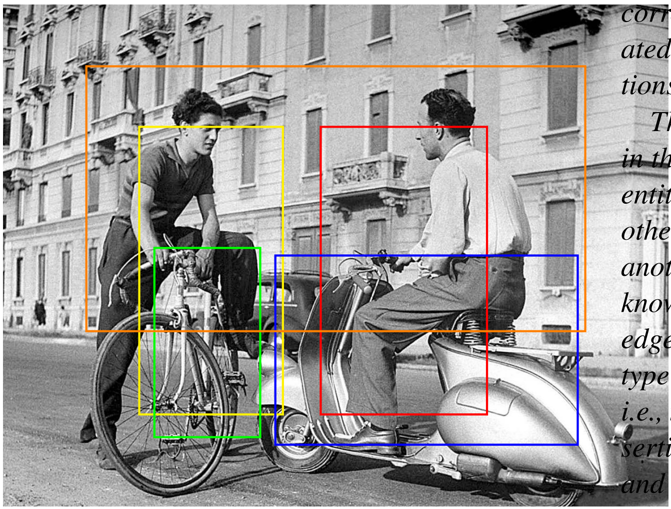

Figure 1: A complex document composed of a picture and a small text, annotated with visual and textual mentions.

Textual-Knowledge Entity Linking (VTKEL) *is the problem of detecting all the entities mentioned in d_t and/or shown in d_i, and linking them to the corresponding named entities in K, if they are present, or linking them to new entities, extending the A-box of K with its type assertion(s), i.e. adding $C(e^{new})$ for each new entity e^{new} of type C mentioned in d.*

Example 2.1 *Consider the document shown in Figure 1, which is composed of one picture and a short sentence (caption) in natural language. As shown in Figure 1, one can find five visual mentions, shown in coloured rectangles in the picture, and five textual mentions, underlined in the text. One could find many more visual mentions in the picture (e.g., windows, road) but let us suppose we are only interested in the mentions of certain types. Let us consider a knowledge base (e.g., YAGO (Suchanek et al., 2007)) that contains knowledge about the named entities e_{Vespa}, e_{Milan}, e_{Italy} and e_{1948} for "Vespa", "Milan", "Italy" and "1948" respectivel, with the respective YAGO types scooter(e_{Vespa}), town(e_{Milan}), country(e_{Italy}), and year(e_{1948}). Let us suppose that the knowledge base contains also the concepts man, and building, and that we focus only on the these mentioned concepts.*

The solution of the VTKEL tasks requires (i) detecting the visual and textual entity mentions of the considered types, and linking them to either (ii) the correct, existing, named entities, or (iii) newly created entities, adding the corresponding type assertions.

The visual and textual mentions of a man shown in the red text and in the red box refer to the same entity, and they should be linked together. The other visual mention of a man, should be linked to another different entity. These two entities are not known (i.e., they are not part of the initial knowledge base K), and therefore two new entities of type man should be added to the knowledge base, i.e., the A-box of K should be extended with the assertions $Man(e^{new}_1)$ and $Man(e^{new}_2)$. The textual and visual mentions of Vespa are also referring to the same entity. However, this time the entity is known (i.e., YAGO contains an entity for Vespa) and therefore the two mentions should be linked to the same entity.[2] *The other visual mentions should be linked to new entities, with the correct type, i.e., we should add the assertions $bicycle(e^{new}_3)$, $building(e^{new}_4)$. For the other textual mentions, i.e., Milan, Italy, 1948, we already have instances in the knowledge base, so we have to link them to these entities.*

VTKEL is a complex task that requires the solution of a set of well studied elementary tasks in natural language processing and computer vision. In particular, the following are the key subtasks of VTKEL:

- (named) entity recognition and classification (i.e. typing) in texts (Goyal et al., 2018);
- object detection in images (Han et al., 2018);
- textual co-reference resolution (Sukthanker et al., 2018);
- textual entity linking to a knowledge base (ontology)(Shen et al., 2015);
- visual entity linking to a knowledge base (ontology) (Tilak et al., 2017);
- visual and textual co-reference resolution (Kong et al., 2014; Huang et al., 2017; Karpathy and Fei-Fei, 2015).

[2] Notice that here, for simplicity we don't take into account the fact that Vespa is a brand and the two mentions are not referring to the brand but to an instance of scooter of this brand. This issue, while surely important, is somehow not central to the main message of the paper, and considering it will make the presentation unnecessarily complex: therefore, we decide to ignore it at the moment.

In the recent literature, a number of approaches focusing on one specific task, or a subset of them, can be found. However, nowadays it is well-established in many areas of NLP and computer vision (see related work) that there is a clear advantage in solving complex tasks in a collective/joint manner, rather than combining the results of task-specific tools used as black-box It is indeed clear that the relation among the appearance of an entity in an image, its associated linguistic properties within the text, and the semantic/axiomatic knowledge contained in the ontology, can jointly contribute to the solution of the complex task altogether. We are particularly interested in pursuing this research direction, and for this reason we need to develop a dataset which is annotated with all the ground truth data needed for the VTKEL problem.

3 Related work

Recently the scientific community of NLP and IP devoted a reasonable effort in investigating the interaction and integration of text and image processing. For a survery of the area of entity information extraction and linking, we refer the reader to (Martinez-Rodriguez et al., 2018), which provides an up-to-date and rather exhaustive survey of the approaches in the area. In particular: (Kong et al., 2014) 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 visual information and visual-textual coreference previously found to solve the coreference in text. The work described in (Weiland et al., 2017, 2018) tackle the problem of ranking the concepts from the knowledge base that best represent 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 (Venkitasubramanian et al., 2017) adopts a statistical model based on Markov Random Fields to represent the dependencies between what is shown in a video frame and its subtitle, and it is applied in the domain of video about wild-life animal. 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 (Miller, 1995). Object detection is not performed: the approach assumes that only one animal is shown in a frame, and the vision part consists in image classification. Furthermore, no background knowledge about animals is used. (Tilak et al., 2017) 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 considered in our work. (Ramanathan et al., 2014) combines textual coreference resolution with the linking of image and textual mentions in order to solve the problem of assigning names of the people of the cast of a TV-show to the tracks of human in the video, leveraging the screenplay scripts accompanying the video as a sort of raw supervision about who's in the video.

Concerning datasets that combine text and images, there are several resources available. However, none of them has all the three components necesary for the VTKEL task. VisualGenome (Krishna et al., 2016) is an extremely large dataset that contains pictures in which objects are annotated along with their types, attributes, and relationships. Each annotation is mapped to some WordNet synset. 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) (Lu et al., 2016) 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 $\langle$object_type, relation, 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 (Lin et al., 2014) 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 relation between object regions, and relation between regions and mentions in the captions. Conceptual Captions (Sharma et al., 2018) 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 no dataset fully annotated with all the informations needed for the VTKEL task. This justifies the development of such a new resource.

4 Background

To build the VTKEL dataset, we start from Flickr30k-Entities dataset (Plummer et al., 2015). Flickr30K Entities provide the annotation of coreference chains, i.e., linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding. They enable us to define a new benchmark for the localization of textual entity mentions in an image. To link textual entities to ontological resource, namely YAGO, we use PIKES (Corcoglioniti et al., 2016). In the following, we summarize these three main components.

Flickr30k-Entities - a dataset with visual and textual mentions alignment The Flickr30k-Entities dataset (Plummer et al., 2015) has become a standard benchmark for sentence-based image description tasks such as image captioning. Flickr30k Entities has augmented the original captions from Flickr30k-Entities with co-reference chains linking mentions of the same entities in images, as well as manually annotated bounding boxes corresponding to each entity. These additional annotations are essential for grounded language understanding of visual data, and they have allowed the recent progress in text-to-image reference resolution and bidirectional image-sentence retrieval. The availability of such ground-truth annotations is also a key resource for experimenting in other high-level tasks, involving both visual and textual data, such as Visual Question Answering (VQA).

YAGO - a large-scale semantic knowledge base YAGO[3] (Suchanek et al., 2007) is a large-scale semantic knowledge base automatically derived from several data sources, including Wikipedia,[4] WordNet,[5] and GeoNames.[6] In particularly, in YAGO an entity (e.g., person, organization, city, etc.) is associated to its corresponding page in Wikipedia, and facts about the entity are extracted from the infobox this Wikipedia page. YAGO entities are typed according to classes organized in a class/sub-class hierarchy obtained combining the categories of Wikipedia with the WordNet synset taxonomy. The current version (v3) of YAGO contains more than 350K classes and 17M entities, with over 150M facts about them.

PIKES - A textual knowledge extraction suite PIKES (Corcoglioniti et al., 2016) is a state-of-the-art frame-based framework for extracting knowledge (graphs) from natural language text. It works in two phases. In the first *linguistic feature extraction* phase, an RDF graph of mentions is obtained by running and combining the outputs of several state-of-the-art NLP tools, including Stanford CoreNLP[7] (tokenization, lemmatization, part-of-speech tagging, temporal expression recognition and normalization, named entity recognition and classification, coreference resolution, parsing), DBpedia Spotlight[8] (entity linking), UKB[9] (word sense disambiguation), Semafor[10] and Mate-tools[11] (semantic role labeling). In the second *knowledge distillation* phase, the mention graph is transformed into an RDF knowledge graph through the evaluation of mapping rules, using the RDFpro[12] (Corcoglioniti et al., 2015) tool for RDF processing. In particular, thanks to the DBpedia-YAGO and WordNet-YAGO mappings, entities resulting in the final RDF knowledge graph are typed according to the classes in YAGO. For the construction of the VTKEL dataset, we run multiple instances of PIKES (using only the minimal set of NLP tools needed for the purpose, namely Stanford CoreNLP, DBpedia

[3] http://yago-knowledge.org

[4] https://www.wikipedia.org/

[5] https://wordnet.princeton.edu/

[6] https://www.geonames.org/

[7] http://nlp.stanford.edu/software/corenlp.shtml

[8] http://spotlight.dbpedia.org/

[9] http://ixa2.si.ehu.es/ukb/

[10] http://www.cs.cmu.edu/~ark/SEMAFOR/

[11] http://code.google.com/p/mate-tools/

[12] http://rdfpro.fbk.eu/

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	10 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

Command	Output
`{\"a}`	ä
`{\^e}`	ê
``{\`i}``	ì
`{\.I}`	İ
`{\o}`	ø
`{\'u}`	ú
`{\aa}`	å

Command	Output
`{\c c}`	ç
`{\u g}`	ğ
`{\l}`	ł
`{\~n}`	ñ
`{\H o}`	ő
`{\v r}`	ř
`{\ss}`	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

Spotlight, and UKB) on a server with 12 cores (24 threads) and 192 GB RAM, obtaining a throughput of ~3500K tokens/h.

5 VTKEL dataset and task

6 Baseline framework for VTKEL

start here:ny full-width figures or tables (see the guidelines in Subsection **??**). **Type single-spaced.** Start all

7 Experiments and Evaluations of VTKEL

start here:ny full-width figures or tables (see the guidelines in Subsection **??**). **Type single-spaced.** Start all

7.1 Sections

Headings: Type

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Using the provided LaTeX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents.

We suggest that instead of

> "(Gusfield, 1997) showed that ..."

you use

> "Gusfield (1997) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

If the BibTeX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LaTeX package. To disable the hyperref package, load the style file with the `nohyperref` option:
`\usepackage[nohyperref]{acl2018}`

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. SIGDAT has **not** adopted the ACL policy of requiring camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier). But we certainly encourage you to use BibTeX records that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`.

As examples, we cite (Goodman et al., 2016) to show you how papers with a DOI will appear in the bibliography. We cite (Harper, 2014) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

Anonymity: As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

> "We previously showed (Gusfield, 1997) ..."

output	natbib	previous SIGDAT style files
(Gusfield, 1997)	`\citep`	`\cite`
Gusfield (1997)	`\citet`	`\newcite`
(1997)	`\citeyearpar`	`\shortcite`

Table 3: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous SIGDAT style files for compatibility.

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

See the `\bibliography` commands near the end for more.

Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

- Example citing an arxiv paper: (Rasooli and Tetreault, 2015).
- Example article in journal citation: (Ando and Zhang, 2005).
- Example article in proceedings, with location: (Borschinger and Johnson, 2011).
- Example article in proceedings, without location: (Andrew and Gao, 2007).

See corresponding .bib file for further details.

7.2 URLs

URLs can be typeset using the `\url` command.

7.3 Footnotes

Footnotes: [13] .[14]

7.4 Graphics

7.5 Accessibility

8 VTKEL baseline framework

It is also.

9 Evaluation

The EMNLP 2019 main conference accepts submissions of long papers and short papers.

(see Appendix A). Papers that d

Conclusion and future work

Acknowledgments

The acknowledgments .

Preparing References:

Include your own bib file like this: `\bibliographystyle{acl_natbib_nourl}` `\bibliography{emnlp2018}`

Where `emnlp2018` corresponds to the `emnlp2018.bib` file.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Benjamin Borschinger and Mark Johnson. 2011. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language*

[13] This is how a footnote should appear.

[14] Note the line separating the footnotes from the text.

Technology Association Workshop 2011, pages 10–18, Canberra, Australia.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. 2016. Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3261–3275.

Francesco Corcoglioniti, Marco Rospocher, Michele Mostarda, and Marco Amadori. 2015. Processing billions of RDF triples on a single machine using streaming and sorting. In *Proc. of ACM Symposium on Applied Computing (SAC)*, pages 368–375. ACM.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11. Association for Computational Linguistics.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. 2018. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the Babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1. Dublin City University and Association for Computational Linguistics.

D. Huang, J. J. Lim, L. Fei-Fei, and J. C. Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1032–1041.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *CVPR*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer.

Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2018. Information extraction meets the semantic web: A survey. *Semantic Web*, pages 1–81. Preprint.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with “their” names using coreference resolution. In *European Conference on Computer Vision*, pages 95–110. Springer.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th International Conference on the World Wide Web*, pages 697–706.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and coreference resolution: A review. *arXiv preprint arXiv:1805.11824*.

Neha Tilak, Sunil Gandhi, and Tim Oates. 2017. Visual entity linking. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 665–672. IEEE.

Aparna Nurani Venkitasubramanian, Tinne Tuytelaars, and Marie-Francine Moens. 2017. Entity linking across vision and language. *Multimedia Tools and Applications*, 76(21):22599–22622.

Lydia Weiland, Ioana Hulpus, Simone Paolo Ponzetto, and Laura Dietz. 2017. Using object detection, nlp, and knowledge bases to understand the message of images. In *International Conference on Multimedia Modeling*, pages 405–418. Springer.

Lydia Weiland, Ioana Hulpuş, Simone Paolo Ponzetto, Wolfgang Effelsberg, and Laura Dietz. 2018. Knowledge-rich image gist understanding beyond literal meaning. *Data & Knowledge Engineering*, 117:114–132.

A Supplemental Material

Each EMNLP 2019 submission can be accompanied by a single PDF appendix, one `.tgz` or `.zip` appendix containing software, and one `.tgz` or `.zip` appendix containing data.

EMNLP 2019 also

Appendices (*i.e.* supplementary material in the form of proofs, tables, or pseudo-code) should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.