Defining Visually Descriptive Language

Robert Gaizauskas[1] **Josiah Wang**[1] **Arnau Ramisa**[2]

[1] Department of Computer Science, University of Sheffield, UK
[2] Institut de Robòtica i Infòrmatica Industrial, Barcelona

{r.gaizauskas, j.k.wang}@sheffield.ac.uk aramisa@iri.upc.edu

Abstract

In this paper, we introduce the notion of *visually descriptive language* (VDL) – intuitively a text segment whose truth can be confirmed by visual sense alone. VDL can be exploited in many vision-based tasks, e.g. image interpretation and story illustration. In contrast to previous work requiring pre-aligned texts and images, we propose a broader definition of VDL that extends to a much larger range of texts without associated images. We also discuss possible VDL annotation tasks and make recommendations for difficult cases. Lastly, we demonstrate the viability of our definition via an annotation exercise across several text genres and analyse inter-annotator agreement. Results show reasonably high levels of agreement between annotators can be reached.

1 Introduction

Recent years have seen rapid growth in research integrating visual and textual modalities, including associating named entities in captions with faces in images (Berg et al., 2004), generating image descriptions (Kulkarni et al., 2011; Yang et al., 2011), text/image retrieval (Hodosh et al., 2013), story illustration (Feng and Lapata, 2010), and learning visual recognition of fine-grained object categories (Wang et al., 2009). This previous work concentrates on solving image-based tasks, and is heavily reliant upon datasets with pre-aligned images and texts, most of which have been manually collected and/or annotated. Thus, such image-centric texts are assumed to be at least partially, if not predominantly, 'visually descriptive' in nature. This raises some interesting research questions: (i) how much text out there without associated images is 'visually descriptive' and thus potentially useful for such image-based tasks? (ii) can these 'visually descriptive' text segments be identified automatically within documents which may consist of predominantly 'non-visual' text?

To be able to answer these questions, we first require a robust, inter-subjectively reliable definition of 'visually descriptive' text. Although previous work exists that models the 'visualness' of terms or concepts from images (Yanai and Barnard, 2005; Jeong et al., 2012), they are presented without an explicit definition apart from the intuitive notion that a visual term should exhibit some consistent visual characteristics across different objects. To our knowledge, the only work that explicitly proposes a definition for visually descriptive text is that of Dodge et al. (2012), where noun phrases within an image caption are classified as to whether or not they are depicted in the corresponding image.

In this paper, we propose a broader definition of *Visually Descriptive Language* (VDL). Our work differs from Dodge et al. (2012) in that our definition revolves around identifying text segments that express propositions that can be 'visually confirmed' rather than identifying 'visually concrete' noun phrase segments whose denotation can be located in an associated image. The consequences of this different definition are significant: (i) we are not restricted to mining VDL from texts with associated images, but can exploit any text, massively extending the volume of data that can be mined; (ii) we can gather larger, richer fragments of text than just noun phrases; (iii) we are not limited to the sort of language found in image captions or texts with embedded images (typically news), but can consider texts of any genre.

It is unlikely there is any one 'correct' definition of VDL. Rather, any proposed definition may be assessed in terms of how useful it is for some particular purpose and how easy it is to apply. Our purpose in defining VDL is to allow us to identify,

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 10–17,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

within a broad corpus of texts, segments that can be used to inform computational models useful in image interpretation and description. For example, co-occurrence in VDL of certain attribute values and object types, or of pairs of objects types, or of object types in particular semantic roles in relation to an activity or event type provide prior information that can be used in Bayesian models to help interpret or describe a new image. Corpora of VDL can also be used to learn language models for generating image descriptions, e.g. for the visually impaired. Other potential applications include identifying candidate text segments within a novel to be illustrated, automatic collection of joint visual-text training data, and automatic extraction of discriminative object descriptions for visual recognition (e.g. butterfly descriptions in Wang et al. (2009)).

1.1 Overview

The rest of the paper is structured as follows. Section 2 presents and discusses our definition of VDL. Section 3 describes possible VDL annotation tasks based on our definition and discusses and makes recommendations on difficult cases. To assess the viability of the definition, we have carried out a pilot annotation exercise on texts of different genres. Section 4 describes and analyses this exercise, including agreement statistics and insights on conflicting annotations. Finally, Section 5 offers conclusions and discusses future work.

2 Definition of VDL

Our intuition is that a segment of text is visually descriptive if we can determine what it says is true or false by visual sense alone. More precisely:

Definition. A text segment is *visually descriptive* iff it asserts one or more propositions about either (a) a specific scene or entity whose truth can be confirmed or disconfirmed through direct visual perception (e.g. (1)), or (b) a class of scenes or entities whose truth with respect to any instance of the class of scenes or entities can be confirmed or disconfirmed through direct visual perception (e.g. (2)).

(1) *John carried the bowl of pasta across the kitchen and placed in on the counter.*

(2) *Tigers have a pattern of dark vertical stripes on reddish-orange fur with a lighter underside.*

(3) * *Maria is thinking about what the future holds for her.* (Not VDL[1])

By *direct visual perception* we mean that:

1. An observer could determine the truth of the relevant proposition without intervening in the scene to acquire additional visual inputs. E.g. the truth of *John weighs 65 kg* might be determined visually by placing John on a scale and taking a reading; but if this scale and Johns standing on it are not part of the scene then this sentence is not VDL.
2. Any inference that needs to be carried out to confirm or disconfirm the proposition is such that it would typically be made by an observer drawn from the population of intended readers of the text without knowledge of the preceding textual content. For example, most observers of a scene that includes a boy sitting on the end of a dock holding a fishing rod whose line disappears into the water before him would infer without question that the boy is fishing, allowing them to confirm the truth of "The boy sat fishing on the dock" directly from the scene and without knowledge of earlier parts of the text in which the sentence is embedded. This example illustrates just how tightly coupled inference and perception are and that "what we see" is a product of both. Also, note how our definition is analogous to that of *textual entailment* where given a pair of textual expressions T and H "We say that T entails H if, typically, a human reading T would infer that H is most likely true" (Dagan et al., 2006); i.e. we rely on a judgement that would *typically* be made about what is going on in the scene.
3. An observer can visually identify any named entities. For example, in (1) we assume an observer knows who John is in this scene. This may only be possible because of knowledge obtained from other textual context, but we don't want to rule out the visualness of (1) on the grounds that not all the information that may be needed to identify John in the scene is present in (1).

By *asserts a proposition* we mean that text segments must express, explicitly or implicitly a predication, i.e. something that may be judged true

[1] We can neither confirm nor disconfirm through direct visual perception alone that Maria is thinking (she might be just staring into space), let alone know what she is thinking.

or false. Sentences or clauses with tensed verbs are candidates, as are noun phrases that predicate something of an entity. Thus, we rule out bare noun phrases (*the man*)[2], but include phrases such as *the tall man* or *a man wearing a green shirt*.

By *text segment* here we mean a phrase, clause, sentence or sequence of sentences, i.e. a sequence of contiguous words. One consequence of this constraint is that phrases like (4) are not VDL, since while they contain a mix of visual (*tall*) and non-visual (*well-educated*) attributes, they do not form a contiguous sequence of words which is visually confirmable as a whole. Since we frequently observed such cases, we want our scheme to accommodate them. We call such segments *impure visually descriptive language* (IVDL). To be IVDL a segment S_1 must contain discontinuous subsequences that if conjoined form a segment S_2 such that (a) S_2 is VDL, and (b) in context S_2 asserts a proposition that is entailed by the proposition S_1 asserts (this rules out conjoining of unrelated subsequences – see (5)). We annotate IVDL subsequences belonging to the same (discontinuous) segment with the same subscript indices.

(4) $\{$*the tall*$\}_1$ *, well-educated* $\{$*man*$\}_1$

Condition (b) serves to rule out cases like:

(5) * $\{$*the tall*$\}_1$ *wardrobe beside the well-educated* $\{$*man*$\}_1$

as from (5) we cannot derive *the tall man*, since the predication it expresses is not entailed by those expressed by (5).

Note that IVDLs are distinct from *partially visual* segments such as (6) containing both visual and non-visual phrasal subcomponents:

(6) *As* $\{$*he walked by the lake*$\}$*, John thought about his dad.*

(6) contains a contiguous sub-segment that is VDL, unlike (4), which is IVDL.

3 Annotating VDL

We describe several possible VDL annotation tasks and provide recommendations on how to approach and annotate some difficult cases.

3.1 Possible Annotation Tasks

We distinguish two annotation tasks: *sentence-level annotation* and *segment-level annotation*.

Sentence-level annotation

We define a sentence-level annotation task as follows. Each sentence S in a document is assigned one of three values: (i) **0** if it contains no VDL; (ii) **1** if the entire sentence is VDL; (iii) **2** if it contains one or more proper sub-segments which are VDL, but the single segment comprising the whole sentence is not VDL. S=**2** may be further classified as **2P** (containing only pure VD sub-segments) and **2I** (contains pure and/or impure VD sub-segments). Variants of the task may be defined depending on whether VDL is taken to include pure VDL only, or to include both pure and impure VDL. In many texts there are significant numbers of impure VDL segments, so omitting them leads to the loss of a substantial quantity of potentially valuable VDL. On the other hand, including them requires substantially more annotation effort and is only likely to be useful if accurate automatic techniques for extracting pure from impure segments can be developed.

Segment-level annotation

Here the exact words comprising a VDL segment are annotated using a swipe and click annotation tool. Variants arise depending on whether one includes impure segments. Note that doing so requires the multiple sequences making up the pure non-contiguous subsequence of the segment to be selected and their association recorded. A simpler, but less informative, alternative is to give the full IVDL segment a distinct code, effectively deferring the task of identifying the pure subsequence in the impure segment. Another variant is to allow annotation to extend over multiple sentences (e.g. to gather action descriptions for interpreting video sequences instead of static scenes). Note that extending the scope of annotation to multiple sentences may affect the content of the annotations. Example (7) is a full single VDL segment in a multi-sentence annotation task.

(7) $\{$*John took a sip of coffee. He read the newspaper for a minute then took a second sip*$\}$

However, as a single-sentence annotation task, the second sentence will be impure as we cannot verify that the sip is a second one (i.e. (8)).

[2] *man* is undoubtedly a visually perceivable entity, but a list of such terms is available under the physical object synset in WordNet and we do not need a programme of text annotation to acquire them.

(8) {*He read the newspaper for a minute then took a*}$_1$ *second* {*sip*}$_1$

Note that sentence-level annotations may be inferred from segment-level annotations.

3.2 Guidelines for difficult cases

Inevitably, various difficult cases emerge during annotation. While it is to be expected that some areas of variation between annotators will unavoidably remain, consistency across annotators is increased and annotation decisions simplified if a standard approach is taken to various anticipated difficult cases. Because of space constraints, here we highlight only a subset of such cases and recommend ways to annotate them. The full set of guidelines, with extensive discussions and examples, is available online[3]. Below we proceed on the assumption that VDL is being annotated at the segment level, sentence-by-sentence.

Metaphors

In general, judgements that *A is like B*, *X appeared to be Y*, *C was as if D* etc., will not be VDL since the judgement of similarity underlying such statements is not something that is likely to be shared by an observer in viewing the entity to which they metaphor is applied. However, the expressions describing the entity to which the metaphor is applied and that supplying the metaphor may themselves be VDL.

(9) *the pews appeared to be* {*broad stairs in a long dungeon*}

(10) *he panted like* {*a big dog that has been running too long*}

Words with mixed visual/aural or visual/experiential meanings

Many words mix visual and aural or visual and experiential senses. For example, verbs like *shout*, *shuffle* and *pant* have an aural and a visual component, not necessarily in the same proportion. Verbs like *shudder* and *flinch*, adjectives like *sombre* and *insolent* (*insolent green eyes*) and adverbs like *deathly* (*deathly pale*) signal not just movement or appearance but also underlying emotional experience or response. Such words should be annotated if visual input alone is judged sufficient to allow a typical observer to unambiguously apply the words, e.g. {*a dreary housing estate*}.

Temporal adverbials of frequency

Temporal adverbs of frequency (*often*, *sometimes*, *usually*) determine how frequently an activity takes place. These are considered VDL, because our imaginary observer could determine visually, over a period of time, how frequently the activity takes place and make an assessment of whether the temporal term applies. The exception is for adverbials that reference calendrical units (*On Tuesdays* {*Bob goes to the park for a picnic*}), because we cannot directly see that it is a Tuesday.

Temporal adverbials of duration

Temporal adverbs of duration determine how long an activity takes. They are marked as VDL where the duration is intuitively assessable as part of the viewing process ({*for a few minutes*}), but not marked when reference to a watch or calendar would be needed for precision or for tracking the extent of the activity (*in 9.58 seconds*, *for two weeks*).

Multiple visual perspectives

Sometimes a sentence may contain information that is visually confirmable, but only from more than one distinct perspective or frame of reference. For example, in (11), an observer could visually confirm that Billy was climbing a tree wearing his backpack. He or she could also visually confirm that the backpack contained various objects. But any position from which an observer could confirm the climbing would not simultaneously allow the visual confirmation of the contents of the backpack.

(11) {*Billy climbed the tree wearing his backpack*}, {*which contained his slingshot, some pebbles and a magnifying glass*}.

In such cases, we advocate annotating distinct VDL segments, one for each visual perspective or frame of reference, as in (11). The reason for this is that we want to derive models of VDL usage that can be used to help interpret or describe images or video that will be taken from a single perspective (at any given time point). Therefore descriptions that mix perspectives are more likely to be confusing than helpful.

Intentional contexts

For the most part, sentences expressing propositional attitudes will not be VDL. However, the sub-constituent that expresses the proposition towards which the speaker has an attitude may well

[3] http://vdlang.github.io/

	Text	Type	|S|	S=1	S=2	VDL	IVDL	%Agree	Kappa	IoU
Oz	Ch7	Children's Story	95	0.13	0.51	51	47	0.76	0.73	0.65
	Ch9	Children's Story	78	0.12	0.42	38	23	0.72	0.69	0.62
Brown	A13	Sports Reportage	111	0.11	0.27	25	20	0.78	0.60	0.51
	A30	Culture Reportage	128	0.04	0.34	31	21	0.78	0.56	0.57
	G32	Biography	101	0.02	0.47	32	29	0.74	0.50	0.43
	L05	Mystery Fiction	151	0.21	0.31	65	20	0.87	0.79	0.63
	N13	Western Fiction	122	0.12	0.46	58	38	0.70	0.49	0.57
	P15	Romance Fiction	179	0.08	0.24	40	21	0.82	0.62	0.73

Table 1: Selected texts and results of the annotation experiment. Column |**S**| shows the number of sentences, columns **S=1** and **S=2** the average proportion of sentences labelled for each VDL type, and columns **VDL** and **IVDL** the number of segments marked as pure and impure VDL on average. Columns **% Agree** and **Kappa** show the inter-annotator agreement at sentence level, and **IoU** the agreement at segment level. Please refer to main text for more details.

be: *John believed that* {*Mary was playing in the garden*}.

Hypotheticals, modals, counterfactuals and subjunctives

Hypothetical or conditional propositions assert something to be the case provided something else is the case. We cannot literally see a conditional, so sentences expressing such propositions are not VDL. However, the antecedent and consequents of such propositions may be visual: *If* {*Jack sets the table*} *then* {*Will serves dinner*}.

Modal (including negation and future tense) and counterfactual sentences may be IVDL since while overall their truth value is not visually determinable, it relates to that of a visually descriptive segment derivable from them. For example, we cannot 'see' that {*James*}$_1$ *may* {*practice Tai Chi in the garden*}$_1$. But the truth of the derived sentence is visually determinable (and is key in possible worlds treatments of the semantics of modals).

Locational information

Locational information is in some cases visually determinable and other cases not. As a general rule any locational information that relies upon geopolitical naming, street plans or compass directions is not marked as VDL. Example (12) is VDL, whist examples (13) and (14) are not VDL.

(12) {*The Episcopal Church stood across the street*}.

(13) *The Episcopal Church was one block down Sussex Street.*

(14) *The Eiffel Tower is in the 7th Arrondissement in Paris.*

Note that although *The Episcopal Church* and *The Eiffel Tower* are named entities and thus visually identifiable according to our definition (see Section 2), locational information may require significant inference using world knowledge that is not part of the text, and thus may not be VDL. For example, we cannot necessarily confirm that someone is in a city called Lisbon based on visual perception alone.

Statements of purpose

Components of sentences that express an agents purpose in doing something should not be annotated as VDL: {*Billy climbed to the rooftop*} *to shoot at crows.*

Imperative and interrogative sentences

Imperative (e.g. (15)) and interrogative (e.g. (16)) sentences do not assert propositions and therefore, by our definition, cannot be VDL as a whole. However, they may contain components which are VDL, for example in (16).

(15) *Come out to the field and call us.*

(16) *How did* {*you escape from the beast*}*?*

Participial phrases

Participial phrases may express predications where they occur within a noun phrase ({*a man wearing a green shirt*}). However, in some cases participial phrases may be extraposed and function, not so much as a reduced relative clause as a sentence adverbial. In this case we annotate across phrasal boundaries, in order to capture the argument of the activity described in the participial phrase, i.e. the entity about which something visual is being predicated. For example in (17), where *John* is included as a VDL segment.

(17) $\{$*Walking slowly across the ice, John*$\}$ thought about his mother.

Dialogues

Text segments that report dialogues do so using either direct (e.g. (18)) or indirect (e.g. (19)) quotation.

(18) *Dorothy said that* $\{$*Toto was running away*$\}$.

(19) *Dorothy said, "*$\{$*Toto is running away*$\}$*".*

In both cases we mark the segment spoken as VDL, if it is VDL. As a matter of convention we do not mark the words reporting who spoken even if we could determine visually whether the person reporting was speaking. This is because (a) these segments are of little interest, and (b) there are many verbs that express fine shades of meaning with respect to spoken utterances, many of which are not visually determinable (*reply*, *ask*, *exhort*, *assert*) and it is easiest just to rule them all out.

4 Results and Analysis

4.1 Experiments and Results

A small pilot annotation exercise was carried out to test the viability of our definition and annotation guidelines on a variety of text genres. As data we used two random chapters from The Wonderful Wizard of Oz and six samples from the Brown Corpus, selected randomly among five hand-picked categories (two news articles, one biography and three novels). As a pilot study, all texts were annotated by the authors at segment-level, the Oz texts by three annotators and the Brown texts by two, using the *brat rapid annotation tool*[4]. Sentence-level annotations are inferred from these segment-level annotations. We chose to annotate at segment level rather than sentence level as identifying VDL segments must be done mentally at sentence level anyway. Marking the segments directly with just a little additional effort will result in a more informative resource.

Table 1 shows the selected texts and an analysis of the resulting annotations. All texts are of similar length (mean 10,834, standard deviation 1,558 characters). Column |**S**| shows the number of sentences in each corpus. Columns **S=1** and **S=2** shows the average proportion of sentences labelled for each VDL type (VDL or partially VDL), and columns **VDL** and **IVDL** the number of *segments* marked as pure and impure VDL on average (rounded to the nearest integer). Percentage agreement (**% Agree**) and **Kappa** are computed at the sentence level. We also report an analysis of the annotation at the segment level: column **IoU** (Intersection-over-Union) shows the overlap of the annotations at word level; i.e. the ratio of words labelled by two annotators as visually descriptive to total number of labelled words by any annotator; at this point we did not distinguish between pure and impure VDL. Figures for the Oz data are averaged pairwise scores over the three annotators.

4.2 Analysis

As Table 1 shows, agreement values are consistently high among annotators and across all genres, supported also by high Kappa scores.

Results show what one would expect: children's stories contain many visual descriptions, hence the higher proportion of VDL sentences and annotator agreement. News articles and biographies contain less VDL than fiction, especially fully visual sentences (column **S=1**). In adult fiction, adventure novels are naturally more visually descriptive than romance, which tends to focus on the mental states and processes of the characters.

Regarding the segment-level analysis, the overlap (**IoU** column) is reasonably high among all texts, indicating that the majority of the visually descriptive phrases were correctly identified. Furthermore, examining the annotations reveals that most inconsistencies are a result of a mistake of just one of the annotators, rather than fundamental difference of opinion, so a revision phase would further increase the agreement.

4.3 Discussion

Further examination of the annotated data revealed some difficult cases in which annotators disagreed. We present and discuss a few example disagreements:

Word with mixed visual/experiential meanings

(20) $\{$*Susan stared at him with hurt blue eyes*$\}_1$.

(21) * $\{$*Susan stared at him with*$\}_1$ *hurt* $\{$*blue eyes*$\}_1$.

Here, *hurt* is used here as an adjective for eyes, which signals both the appearance of the eyes and

[4] http://brat.nlplab.org/

an underlying emotion within Susan. We believe that *hurt* can be accurately applied based on the appearance of Susan's eyes alone, and thus include it as part of the VDL segment.

Inference

(22) {*Rourke was talking on the phone when he came*}$_1$ *back.*

(23) * {*Rourke was*}$_1$ *talking* {*on the phone when he came*}$_1$ *back.*

As with the fishing example in Section 2, most observers may infer that Rourke is talking on the phone from a scene that involves him holding a phone by his ear while moving his mouth. Thus, we consider *talking on the phone* in this context as VDL.

Context

(24) {*The Lion went back*}$_1$ *a third time* {*and got the Tin Woodman*}$_1$.

(25) * {*The Lion went back a third time and got the Tin Woodman*}.

Without context, the annotator has no knowledge about the previous two attempts. Therefore, *a third time* is considered not VDL.

Visual observations over long periods

(26) *From the way* {*the wound in his head*} *was itching, Dan knew that it would heal.*

(27) * *From the way* {*the wound in his head*} *was itching, Dan knew that* {*it*}$_1$ *would* {*heal*}$_1$.

Although one is able to observe a wound healing, it is a very slow process that spans a long period, analogous to watching grass grow. To be able to confirm this proposition would require observation over a long period of time. Therefore, it is preferable not to annotate such cases as VDL.

Explicit naming of entities

(28) {*They go to school with a girl*} *named Gloriana*

(29) * {*They go to school with a girl named Gloriana*}

According to our definition an observer can visually identify any named entities. However, in this particular case, we cannot visually confirm that the name of the girl is Gloriana. Contrast this to {*They go to school with Gloriana*}, where Gloriana is a known named entity, and we can visually confirm the proposition asserted by the text segment.

Directional information

(30) {*He crossed the street and walked swiftly*}$_1$ *southward* {*to circle back to the Boulevard and*}$_1$ *north* {*a block to the open restaurant.*}$_1$

(31) * {*He crossed the street and walked swiftly southward to circle back to the Boulevard and north a block to the open restaurant.*}

It is stated in the guidelines that locational information that relies on compass directions should not be marked as VDL. Example (30) is thus the correct annotation.

Subjective opinions

(32) *They must be* {*dreadful beasts*}.

(33) * {*They*}$_1$ *must* {*be*}$_1$ *dreadful* {*beasts*}$_1$.

(34) * *They must be dreadful beasts.*

Here, the adjective *dreadful* should be considered VDL if it would typically be inferred by an observer given only visual input. Clearly *dreadful* also has an experiential sense, dependent on the subjective impression made on the observer. So the question is are a vast majority likely to agree that the beasts are dreadful? In this case, we accept (32) as valid, although a more complete version would be {*They*}$_1$ *must* {*be dreadful beasts*}$_1$.

Intensifier adverbials and Negation of entities

(35) {*The sides were so steep*} *that* {*none of them*}$_1$ *could* {*climb down*}$_1$

(36) {*The sides were*}$_1$ *so* {*steep*}$_1$ *that none of them could climb down*

(37) {*The sides were so steep that none of them could climb down*}

This is a difficult case where all three annotators annotated differently. Our guidelines did not address cases of adverbs such as *so* and *too*. Another issue that was not addressed in the guidelines is how to deal with the negation of entities (*none of them*). We hope to address these issues in future iterations of the guidelines.

5 Conclusion and future work

In this work we have offered a precise definition of Visually Descriptive Language (VDL), a notion with many possible applications at the intersection of language and vision, a subject of increasing interest. We have conducted a pilot annotation exercise, showing that the proposed definition and annotation guidelines can be used to successfully identify visual fragments in documents of different genres with good levels of agreement across annotators.

We believe that VDL is a useful concept to further stimulate research integrating language and vision. In the future we aim to further refine the proposed annotation guidelines, to explore the feasibility of adapting the annotation task for large-scale crowd-sourcing and to extract features and train models for automatically detecting visual fragments in new documents.

Acknowledgments

This work was funded by the ERA-net CHIST-ERA D2K 2011 VisualSense project (UK EP-SRC EP/K019082/1 and Spanish MINECO PCIN-2013-047). This work was also partly funded by the Spanish MINECO project RobInstruct TIN2014-58178-R.

References

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A. Forsyth. 2004. Names and faces in the news. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pages 848–854.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Proceedings of the First international conference on Machine Learning Challenges: evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, 47:853–899.

Jin-Woo Jeong, Xin-Jing Wang, and Dong-Ho Lee. 2012. Towards measuring the visualness of a concept. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2415–2418.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Josiah Wang, Katja Markert, and Mark Everingham. 2009. Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference*.

Keiji Yanai and Kobus Barnard. 2005. Image region entropy: A measure of "visualness" of web images associated with one concept. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 419–422.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454. Association for Computational Linguistics.