Defining Visually Descriptive Language

Robert Gaizauskas[§], Josiah Wang[§], Arnau Ramisa[¶]

[§] *Department of Computer Science, University of Sheffield, UK*
[¶] *Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Barcelona, Spain*

http://vdlang.github.io

Introduction

We introduce ***Visually Descriptive Language*** (VDL) — intuitively, a text segment whose truth can be confirmed by visual sense alone.

He shot through the air and landed safely on the other side. They were all greatly pleased to see how easily he did it, and after the Scarecrow had got down from his back the Lion sprang across the ditch again. Dorothy thought she would go next; so she took Toto in her arms and climbed on the Lion's back, holding tightly to his mane with one hand. The next moment it seemed as if she were flying through the air; and then, before she had time to think about it, she was safe on the other side. The Lion went back a third time and got the Tin Woodman, and then they all sat down for a few moments to give the beast a chance to rest.

Definition of Visually Descriptive Language (VDL)

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

John carried the bowl of pasta across the kitchen and placed it on the counter.

(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

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

The example below is ***not*** VDL: you cannot confirm that Maria is thinking (she might just be staring into space!)

✗ Maria is thinking about what the future holds for her.

VDL need not be a complete sentence. A VDL segment may be part of a sentence:

As he walked by the lake, John thought about his dad.

Impure VDL (IVDL): Discontiguous subsequences if conjoined may form VDL, but only if entailed by the original proposition.

The tall, well-educated man

Why?

Dodge et al. (2012) classify noun phrases in captions as to whether they are depicted in the corresponding image

Our definition works on any text (no images needed)

We cover larger fragments of texts beyond noun phrases, and can consider any genre

Applications

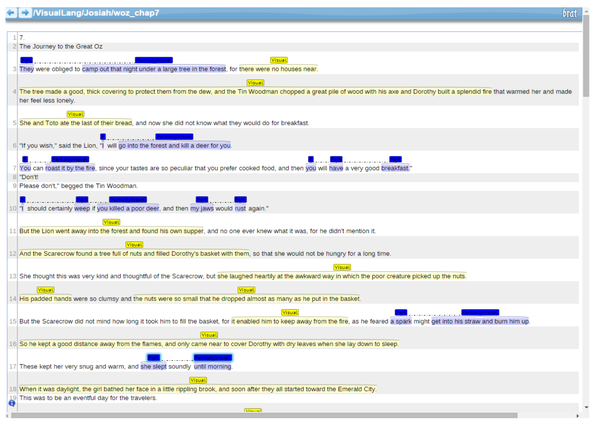
Compute co-occurrence priors (object pairs, object-attributes, etc.)

Learn language models for generating image descriptions

Detect text segments for illustrating novels

Annotation

The Wonderful Wizard of Oz (2 chapters) (3 annotators)
Brown Corpus (6 samples from 5 categories) (2 annotators)



Agreement

Sentence-level annotation inferred from segment-level

	Text	Type	|S|	S=V	S=PV	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

* V = Visual, PV = Partially Visual, IoU = Word-level overlap (Intersection over Union)

Reasonably high agreement and overlap
Children's Story & Fiction: More VDL than News & Biography
Mystery & Western Fiction: More VDL than Romance

Future Work

Refine annotation guidelines
Adapt annotation task for crowd-sourcing
Train models for detecting VDL in new documents

Examples and Difficult Cases

Mixed visual/experiential meanings: A *dreary* housing estate.
Mixed visual/aural meanings: He *shouted* at the waiter.

} VDL if can be unambiguously applied by typical observer using visual sense alone

Temporal adverbials of frequency are VDL, unless a calendar is needed:
Bob *often* goes to the park for a picnic.
On Tuesdays, Bob goes to the park for a picnic.

Temporal adverbials of duration are VDL, unless a watch/calendar is needed for precision/tracking:
He stopped for *a few minutes*.
She slept for *9.58 minutes*.

Metaphors: Generally not VDL, but the expression supplying the metaphor may be VDL:
The pews appeared to be broad stairs in a long dungeon.

Intentional context:
Sarah thought that Molly was playing in the garden.

Hypotheticals:
If Jack sets the table then Will serves dinner.

Modals:
James may practice Tai Chi in the garden.

Example Disagreements

✓ Rourke was talking on the phone when he came back.
✗ Rourke was talking on the phone when he came back.

} One would typically be able to infer that Rourke is talking on the phone from a scene with him holding a phone while moving his mouth.

Without context, the annotator has no knowledge about the previous two attempts.

{ The lion went back a third time and got the Tin Woodman. ✓
{ The lion went back a third time and got the Tin Woodman. ✗

Funding Acknowledgments

chist-era Visual Sense project (EPSRC EP/K019082/1, MINECO PCIN-2013-047) & MINECO RobInstruct Project (TIN2014-58178-R).

The University Of Sheffield.

Institut de Robòtica i Informàtica Industrial