

Phrase Localization Without Paired Training Examples (Supplementary document)

Josiah Wang
Imperial College London
<http://www.josiahwang.com>

Lucia Specia
Imperial College London
l.specia@imperial.ac.uk

Appendix

Figure 1 gives a full-sized illustration of our proposed model for phrase localization without paired examples.

Table 1 shows the complete phrase localization accuracies on the Flickr30kEntities dataset, for different detector combinations, semantic similarities and localization strategies. The upper-bounds for each are also provided (the best possible achievable accuracy before performing the phrase localization step). The upper-bound variant **+union** includes the minimal bounding box encompassing all candidates as a candidate bounding box, while **-union** does not. Note that the **places365** detector is equivalent to a baseline that always localizes a phrase to the whole image.

Table 2 shows the complete accuracies on the ReferItGame test dataset, again for different combinations of detectors, concept selection measure and localization strategies.

Figures 2 and 3 show some example output from our model on Flickr30kEntities and ReferItGame respectively, where we compare the effects of combining a **tfoid** detector with a **tf coco** detector.

Figures 4 and 5 show some example output from our model on Flickr30kEntities and ReferItGame respectively, where we compare the effects of combining a **colour** detector with a **tf coco+tfoid+places365** detector.

Figures 6 and 7 show some example output from our model on Flickr30kEntities and ReferItGame respectively, where we compare the effects of using our proposed **consensus** localization strategy against a **union** strategy.

Detector	Similarity	Upperbound		Localization strategy				
		-union	+union	random	largest	confidence	union	consensus
tfcoco20	no filter	51.67	58.83	14.62	32.89	30.23	30.93	-
	w2v-max	45.28	51.54	22.21	35.98	32.62	36.33	-
	w2v-last	44.92	51.16	21.52	35.72	32.40	36.13	-
	w2v-avg	45.49	51.81	22.26	36.13	32.79	36.49	-
tfcoco	no filter	58.22	65.31	11.37	32.53	30.08	29.81	-
	w2v-max	45.32	51.22	24.10	37.03	34.20	37.57	-
	w2v-last	45.30	51.21	24.51	37.03	34.18	37.70	-
	w2v-avg	45.17	51.29	24.06	37.03	34.11	37.81	-
tfoid	no filter	54.26	63.26	17.28	28.95	21.86	30.90	-
	w2v-max	43.83	50.04	38.27	40.59	39.77	44.69	-
	w2v-last	43.69	49.98	38.30	40.47	39.69	44.54	-
	w2v-avg	43.24	49.56	37.86	40.04	39.22	44.31	-
yolo9000	no filter	24.94	36.25	6.97	18.20	11.82	22.37	-
	w2v-max	15.14	20.85	11.82	13.74	12.51	17.79	-
	w2v-last	15.10	20.69	11.53	13.76	12.57	17.72	-
	w2v-avg	15.14	20.99	11.50	13.76	12.53	17.89	-
places365	no filter	21.99	21.99	21.99	21.99	21.99	21.99	-
	w2v-max	21.99	21.99	21.99	21.99	21.99	21.99	-
	w2v-last	21.99	21.99	21.99	21.99	21.99	21.99	-
	w2v-avg	21.99	21.99	21.99	21.99	21.99	21.99	-
color	no filter	31.71	43.64	1.06	13.91	4.32	22.26	-
	w2v-max	10.19	23.11	3.67	8.51	5.13	18.51	-
	w2v-last	6.53	18.67	2.23	5.50	2.88	15.87	-
	w2v-avg	10.26	23.23	3.85	8.62	5.19	18.65	-
tfcoco+tfoid	no filter	67.24	73.00	13.39	30.32	30.32	27.47	-
	w2v-max	48.84	55.85	38.62	44.00	41.95	48.20	47.91
	w2v-last	48.70	55.80	38.63	43.92	41.83	48.14	-
	w2v-avg	48.31	55.37	38.21	43.53	41.44	47.86	47.63
tfcoco+tfoid+yolo9000	no filter	67.98	73.20	11.77	28.42	30.32	23.09	-
	w2v-max	46.68	53.34	37.38	42.41	40.70	46.36	46.03
	w2v-last	46.76	53.44	37.43	42.55	40.81	46.36	-
	w2v-avg	45.22	51.99	36.58	41.21	39.74	45.37	45.22
tfcoco+tfoid+places365	no filter	72.63	73.37	13.85	22.97	30.34	22.73	-
	w2v-max	51.56	57.81	41.70	46.96	45.00	50.49	50.29
	w2v-last	51.45	57.81	41.78	46.87	44.90	50.46	-
	w2v-avg	51.21	57.51	41.70	46.63	44.71	50.27	50.11
tfcoco+tfoid+places365+yolo9000	no filter	73.31	73.99	12.01	22.79	30.34	19.12	-
	w2v-max	49.52	55.59	40.71	45.35	43.74	48.73	48.51
	w2v-last	49.57	55.71	40.73	45.40	43.77	48.79	-
	w2v-avg	48.17	54.36	39.81	44.20	42.86	47.84	47.79
tfcoco+tfoid+places365+color	no filter	78.64	78.98	5.45	22.51	30.34	21.96	-
	w2v-max	50.24	57.88	38.19	45.75	43.09	48.69	48.91
	w2v-last	50.93	57.81	41.04	46.41	44.13	50.36	-
	w2v-avg	50.47	58.10	39.33	45.78	42.93	49.61	49.51
tfcoco+tfoid+places365+yolo9000+color	no filter	79.10	79.44	5.69	22.53	30.34	18.63	-
	w2v-max	48.59	56.05	37.09	44.44	42.10	47.21	47.45
	w2v-last	49.17	55.81	39.97	45.00	43.08	48.74	-
	w2v-avg	48.01	55.53	37.86	43.81	41.52	47.63	47.61

Table 1. Non-paired phrase localization accuracies for Flickr30kEntities.

Detector	Similarity	Upperbound		Localization strategy				
		-union	+union	random	largest	confidence	union	consensus
tfcoco20	no filter	27.21	30.33	8.22	13.20	11.93	12.91	-
	w2v-max	24.35	26.82	11.01	14.97	14.31	13.97	-
	w2v-last	19.48	21.82	9.83	13.26	12.39	13.26	-
	w2v-avg	24.20	26.64	10.77	14.92	14.24	13.92	-
tfcoco	no filter	31.84	35.18	7.15	13.35	11.79	13.00	-
	w2v-max	25.00	27.16	11.49	15.40	14.73	13.85	-
	w2v-last	17.13	19.08	9.55	12.57	11.73	12.51	-
	w2v-avg	24.43	26.54	11.25	15.24	14.55	13.73	-
tfoid	no filter	32.75	36.85	10.55	14.82	12.93	14.61	-
	w2v-max	25.46	28.03	18.48	19.82	19.87	18.88	-
	w2v-last	19.98	22.36	15.23	16.49	16.24	16.52	-
	w2v-avg	25.13	27.73	18.35	19.68	19.71	18.89	-
yolo9000	no filter	9.75	13.22	6.71	8.37	7.09	10.34	-
	w2v-max	7.93	9.56	7.32	7.64	7.45	8.95	-
	w2v-last	7.76	9.30	7.24	7.50	7.34	8.78	-
	w2v-avg	7.93	9.57	7.32	7.63	7.44	8.95	-
places365	no filter	14.64	14.64	14.64	14.64	14.64	14.64	-
	w2v-max	14.64	14.64	14.64	14.64	14.64	14.64	-
	w2v-last	14.64	14.64	14.64	14.64	14.64	14.64	-
	w2v-avg	14.64	14.64	14.64	14.64	14.64	14.64	-
color	no filter	40.43	45.83	2.04	14.74	7.92	14.52	-
	w2v-max	13.92	20.52	5.82	11.08	8.49	13.99	-
	w2v-last	10.69	16.85	4.70	8.78	6.79	12.41	-
	w2v-avg	13.82	20.58	5.93	11.05	8.61	14.10	-
tfcoco+tfoid	no filter	42.73	46.33	7.76	15.00	12.36	14.39	-
	w2v-max	30.35	33.10	18.06	21.21	20.69	19.45	20.10
	w2v-last	20.95	23.39	13.77	16.53	15.49	16.44	-
	w2v-avg	29.99	32.70	17.78	21.21	20.55	19.38	19.68
tfcoco+tfoid+yolo9000	no filter	43.00	46.63	7.13	14.01	12.24	13.15	-
	w2v-max	28.44	31.13	17.04	19.88	19.54	18.40	19.30
	w2v-last	18.87	21.29	12.80	15.21	14.35	15.48	-
	w2v-avg	26.99	29.70	16.56	19.46	19.15	18.14	18.53
tfcoco+tfoid+places365	no filter	48.80	49.14	8.76	14.65	12.53	14.58	-
	w2v-max	33.00	35.45	20.85	23.93	23.57	21.97	22.64
	w2v-last	22.84	24.98	16.40	18.81	17.85	18.67	-
	w2v-avg	32.61	35.04	20.93	23.95	23.43	21.95	22.25
tfcoco+tfoid+places365+yolo9000	no filter	49.08	49.42	8.03	14.53	12.50	13.52	-
	w2v-max	31.07	33.55	19.77	22.55	22.34	20.94	21.83
	w2v-last	20.98	23.16	15.41	17.57	16.77	17.76	-
	w2v-avg	29.53	32.03	19.50	22.08	21.87	20.64	21.01
tfcoco+tfoid+places365+color	no filter	69.78	70.00	4.21	14.81	12.58	14.49	-
	w2v-max	35.57	39.50	21.22	26.48	25.36	24.52	25.52
	w2v-last	23.82	27.24	15.95	20.05	18.36	20.49	-
	w2v-avg	34.91	38.97	20.96	26.45	24.92	24.66	25.11
tfcoco+tfoid+places365+yolo9000+color	no filter	69.93	70.13	4.02	14.78	12.58	13.43	-
	w2v-max	34.19	38.07	20.45	25.44	24.51	23.69	24.86
	w2v-last	22.76	26.13	15.57	19.23	17.76	19.82	-
	w2v-avg	32.51	36.52	20.01	24.98	23.78	23.57	24.08

Table 2. Non-paired phrase localization accuracies for ReferItGame.

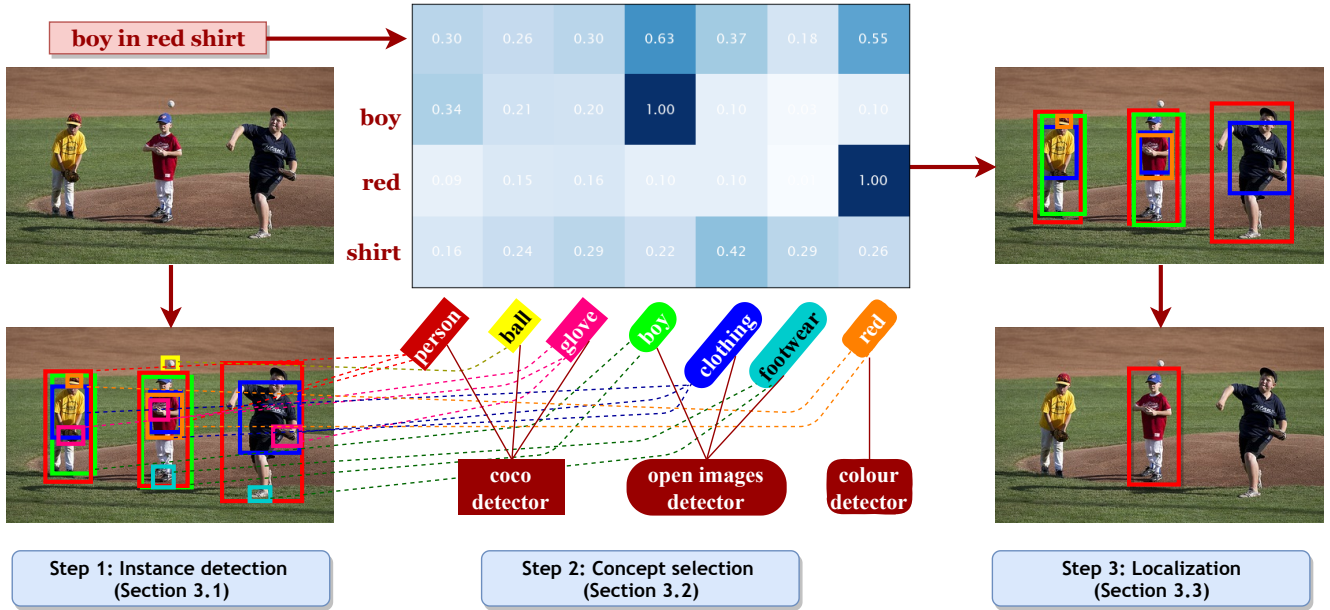


Figure 1. The three stages of the proposed model for unsupervised phrase localization. The **instance detection** phase detects instances of various concepts using pre-trained detectors. The **concept selection** stage ranks these detected concepts against the query phrase (using pre-trained word embeddings) and forwards the best candidate concept instance(s) to the **localization** phase, where the model predicts the final bounding box for the query phrase.



Figure 2. Example localization output for Flickr30kEntities. We compare the effects of adding a **tfoid** detector (red bounding box) to **tfcoco** (blue) (**w2v-max**, **union**). The ground truth is indicated in green. The first row shows examples of where adding a **tfoid** detector improves localization, while the last row provides examples where it hurts localization.

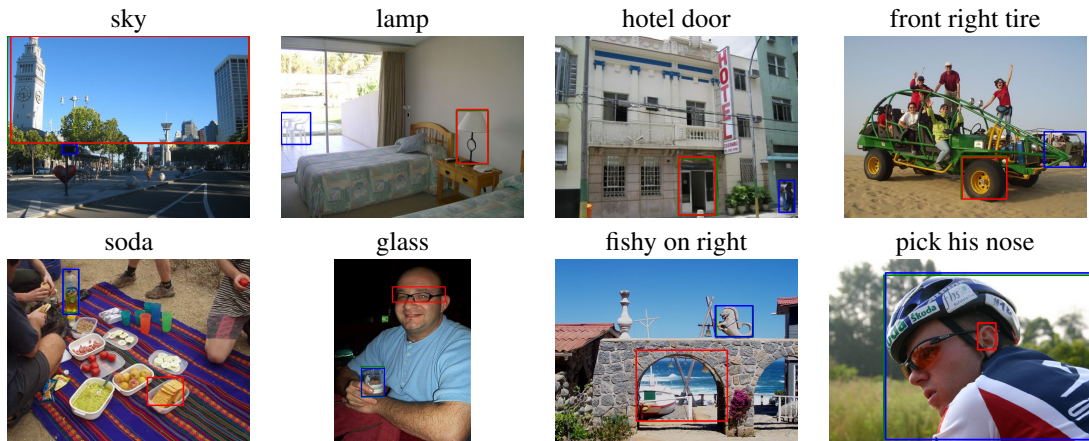


Figure 3. Example localization output for ReferItGame. We compare the effects of adding a **tfoid** detector (red bounding box) to **tf coco** (blue) (**w2v-max**, **union**). The ground truth is indicated in green. The first row shows examples of where adding a **tfoid** detector improves localization, while the last row provides examples where it hurts localization.

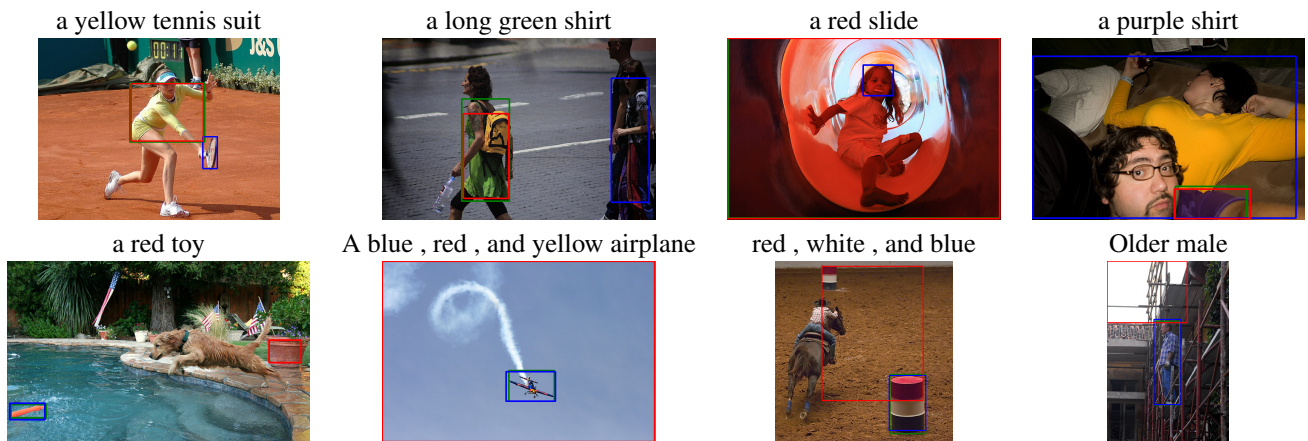


Figure 4. Example localization output for Flickr30kEntities. We compare the effects of adding a **colour** detector (red bounding box) to **tf coco+tfoid+places365** (blue) (**w2v-avg**, **union**). The ground truth is indicated in green. The first row shows examples of where adding a **colour** detector improves localization, while the last row provides examples where it hurts localization.



Figure 5. Example localization output for ReferItGame. We compare the effects of adding a **colour** detector (red bounding box) to **tfccoco+tfoid+places365** (blue) (**w2v-avg, union**). The ground truth is indicated in green. The first row shows examples of where adding a **colour** detector improves localization, while the last row provides examples where it hurts localization.

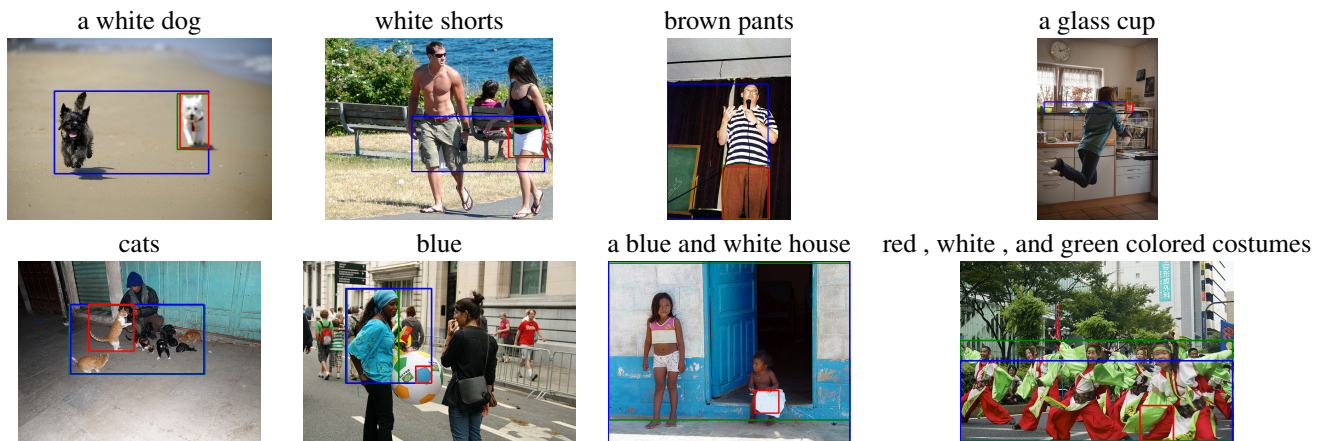


Figure 6. Example localization output for Flickr30kEntities. We compare the effects of using a **consensus** (red bounding box) against a **union** (blue) localization strategy (with **tfccoco+tfoid+places365+color, w2v-avg**). The ground truth is indicated in green. The first row shows examples of where using **consensus** improves localization, while the last row provides examples where it hurts localization.

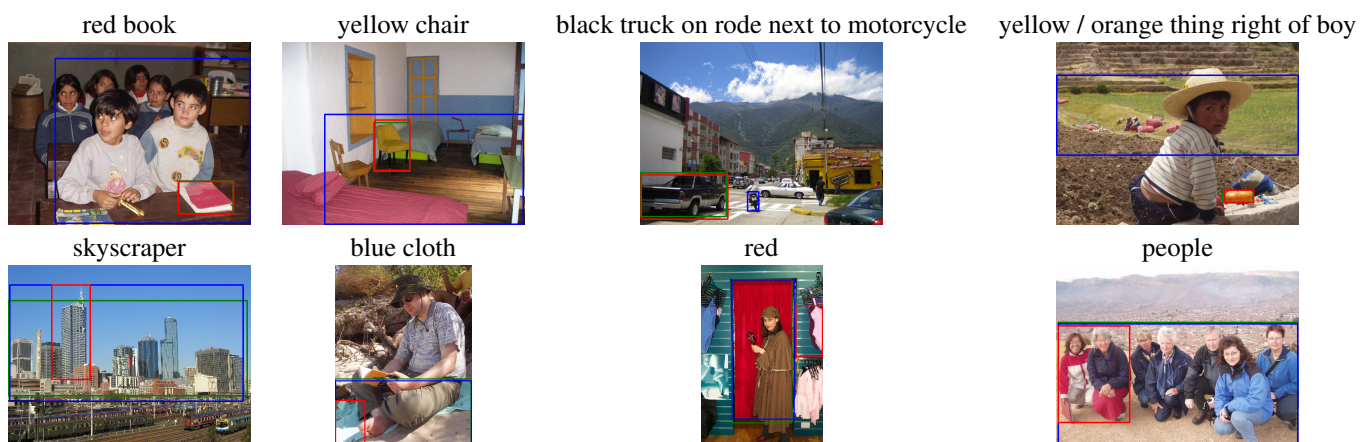


Figure 7. Example localization output for ReferItGame. We compare the effects of using a **consensus** (red bounding box) against a **union** (blue) localization strategy (with **tf coco+tfoid+places365+color, w2v-avg**). The ground truth is indicated in green. The first row shows examples of where using **consensus** improves localization, while the last row provides examples where it hurts localization.