A Semantic evaluation results table

This document is a supplement to our paper titled: Learning semantic sentence representations from visually grounded language without lexical knowledge (Merkx and Frank, 2019). It contains the semantic evaluation results which were used to create Figure 3 of said paper.

Table 1: Semantic textual similarity results (Pearson's r ×100 with 95 percent confidence interval). BOW is a bag of words approach using GloVe embeddings and InferSent is the model reported by Conneau et al. (2017)

		Model			
Task	Dataset	BOW	InferSent	char-GRU	char-GRU
				(Flickr8k)	(MSCOCO)
STS 2012	MSRpar	$42.3^{+5.7}_{-6.1}$	$40.0^{+5.8}_{-6.2}$	$49.1^{+5.2}_{-5.6}$	$37.6^{+6.0}_{-6.3}$
	MSRvid	$66.2^{+3.8}_{-4.2}$	$83.6^{+2.0}_{-2.3}$	$79.9^{+2.4}_{-2.7}$	$82.7^{+2.1}_{-2.4}$
	SMTeuroparl	$48.4^{+6.7}_{-7.3}$	$47.1^{+6.8}_{-7.4}$	$57.3^{+5.8}_{-6.5}$	$54.2^{+6.2}_{-6.8}$
	OnWN	$57.0^{+4.6}_{-5.0}$	$64.5^{+4.0}_{-4.4}$	$67.5^{+3.7}_{-4.1}$	$65.5^{+3.9}_{-4.3}$
	SMTnews	$46.3^{+7.4}_{-8.1}$	$60.7^{+5.9}_{-6.6}$	$51.1^{+6.9}_{-7.6}$	$38.7^{+8.0}_{-8.7}$
STS 2013	FNWN	$38.2^{+11.6}_{-12.9}$	$34.5^{+12.0}_{-13.2}$	$23.8^{+13.0}_{-13.9}$	$23.2^{+13.1}_{-14.0}$
	HDL	$63.4^{+4.1}_{-4.5}$	$69.0^{+3.6}_{-3.9}$	$67.3^{+3.7}_{-4.1}$	$64.9^{+4.0}_{-4.3}$
	OnWN	$47.2^{+6.2}_{-6.7}$	$73.1^{+3.6}_{-4.1}$	$59.8^{+5.1}_{-5.6}$	$58.5^{+5.2}_{-5.7}$
STS 2014	Deft-forum	$30.0^{+8.2}_{-8.7}$	$47.5^{+6.9}_{-7.5}$	$50.7^{+6.6}_{-7.2}$	$51.5^{+6.5}_{-7.1}$
	Deft-news	$65.0^{+6.1}_{-7.1}$	$72.9^{+4.9}_{-5.8}$	$67.8^{+5.7}_{-6.6}$	$65.3^{+6.0}_{-7.0}$
	HDL	$58.7^{+4.5}_{-4.9}$	$63.6^{+4.1}_{-4.5}$	$61.6^{+4.3}_{-4.6}$	$60.0^{+4.4}_{-4.8}$
	Images	$62.4^{+4.2}_{-4.6}$	$80.9^{+2.3}_{-2.6}$	$81.4^{+2.3}_{-2.6}$	$88.2^{+1.5}_{-1.7}$
	OnWN	$57.7^{+4.6}_{-5.0}$	$77.3^{+2.7}_{-3.1}$	$68.6^{+3.6}_{-4.0}$	$68.1^{+3.7}_{-4.0}$
	Tweet-news	$53.9^{+4.9}_{-5.3}$	$75.3^{+2.9}_{-3.3}$	$74.0^{+3.1}_{-3.4}$	$69.6^{+3.5}_{-3.9}$
STS 2015	Answers forum	$36.7^{+8.4}_{-9.1}$	$61.3^{+6.0}_{-6.7}$	$57.6^{+6.4}_{-7.2}$	$49.4^{+7.3}_{-8.1}$
	Answers student	$63.6^{+4.1}_{-4.5}$	$68.6^{+3.6}_{-4.0}$	$68.8^{+3.6}_{-4.0}$	$67.1^{+3.8}_{-4.1}$
	belief	$44.8^{+7.7}_{-8.5}$	$71.8^{+4.6}_{-5.3}$	$71.8^{+4.6}_{-5.3}$	$66.5^{+5.3}_{-6.1}$
	HDL	$66.2^{+3.8}_{-4.2}$	$69.6^{+3.5}_{-3.9}$	$70.3^{+3.4}_{-3.8}$	$67.1^{+3.8}_{-4.1}$
	Images	$69.1^{+3.6}_{-3.9}$	$85.5^{+1.8}_{-2.1}$	$89.2^{+1.4}_{-1.6}$	$88.1^{+1.5}_{-1.7}$
STS 2016	Answer-Answer	$40.1^{+9.8}_{-10.9}$	$62.0^{+7.0}_{-8.2}$	$53.8^{+8.2}_{-9.4}$	$47.8^{+9.0}_{-10.1}$
	HDL	$61.4^{+7.2}_{-8.4}$	$68.8^{+6.0}_{-7.2}$	$70.0^{+5.8}_{-6.9}$	$65.4^{+6.6}_{-7.7}$
	Plagiarism	$54.6^{+8.5}_{-9.8}$	$80.8^{+4.1}_{-5.0}$	$78.6^{+4.5}_{-5.5}$	$75.8^{+5.0}_{-6.1}$
	Postediting	$53.9^{+8.3}_{-9.6}$	$82.3^{+3.7}_{-4.5}$	$84.4^{+3.3}_{-4.0}$	$79.7^{+4.2}_{-5.1}$
	Question-Question	$47.2^{+9.9}_{-11.3}$	$63.3^{+7.5}_{-8.9}$	$40.5^{+10.8}_{-12.0}$	$49.9^{+9.5}_{-10.9}$
STS-B	STS 12-16	$64.7^{+3.0}_{-3.2}$	$75.7^{+2.2}_{-2.3}$	$72.2^{+2.4}_{-2.6}$	$70.7^{+2.5}_{-2.7}$
SICK	Relatedness	$79.9^{+1.0}_{-1.0}$	$86.2^{+0.7}_{-0.7}$	$81.5^{+0.9}_{-1.0}$	$82.7^{+0.9}_{-0.9}$

References

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP*. ACL, 2017.

Danny Merkx and Stefan Frank. Learning semantic sentence representations from visually grounded language without lexical knowledge. *NLE, special issue on sentence representations*, 2019.