Appendix

A Error Analysis

While the proposed approaches enjoy strong performance on the STS benchmarks relative to the competing methods, the Pearson correlations between gold and system scores remain consistently below 0.9 in all subtasks. It would be extremely useful to establish which similarities are not captured very well by these approaches, at least as judged by humans on the 0 to 5 scale established in (Agirre et al., 2012). For concreteness, we limit our exposition to MaxPool Spearman, noting that similar conclusions hold for CKA-based methods too.

First, we linearly transform the system scores into the range $[0, 5]$, thus making them comparable to gold scores while preserving Pearson correlation. Then, in each subtask we select 5 sentence pairs with the largest absolute difference between the gold and the system score. After that, we manually examine the obtained dataset, focusing predominantly on shorter sentences, where the errors are often obvious and easy to explain. Even under these restrictions, we can readily distinguish between 5 different types of errors, summarised in Table 1. On the one hand, the system heavily underestimates the similarity score when two sentences use completely different vocabulary yet have identical meaning (Type I). On the other hand, it tends to overestimate the similarity when the sentences use very related or even the same words but have different meaning (Types II & III). The similarity is also overestimated when two sentences contain antonyms, or when one sentence is a negation of the other (Types IV & V respectively). A lot of these flaws can be traced back to the well-known weaknesses of word embeddings and the distributional hypothesis, such as mixing together semantic similarity and conceptual relatedness (Hill et al., 2015; Mrkšić et al., 2016), failure to distinguish synonyms from antonyms (Mohammad et al., 2008; Mrkšić et al., 2016) and problems with negation. We hope that any counter-measures to these weaknesses will also improve the proposed sentence-level systems.

B Significance Analysis

Following the procedure described in Zhelezniak et al. (2019), we construct 95% BCa confidence intervals for the delta in performance between two systems. The key results are as follows. MaxPool Spearman overall statistically outperforms DynaMax Zhelezniak et al. (2019) when word vectors are highly non-normal (GloVe) and looses when word vectors seem mostly normal (word2vec), which is in line with our main discussion. Next, max-pooling outperforms mean-pooling on the majority of subtasks for all word vector models. Finally, MaxPool Spearman is overall comparable to CKA Gaussian, with the exception of word2vec where CKA is slightly better.

Type	Sentence 1	Sentence 2	Gold	Sys.	Δ
I	*Identical meaning but different words*				
	restrict or confine	place limits on (extent or access).	4.75	1.69	+3.06
	the reduction of the extent of something, e.g, its size, importance or quantity	change toward something smaller or lower.	4.4	1.49	+2.91
	an occasion on which people can assemble for social interaction and entertainment.	festive social event, celebration	4.25	1.33	+2.92
II	*Related words but different meaning*				
	a man is playing the piano.	a woman is playing the violin.	1	3.66	-2.66
	indonesian president to visit uk	indonesian president to visit australia	1.4	4.29	-2.89
	a grey, black, and white cat looking at the camera.	a black and white dog looking at the camera.	1	3.95	-2.95
III	*Same keywords but different meaning*				
	why do you need to peel peaches to can them?	how to peel peaches?	1	4.57	-3.57
	what does it mean to write a song in a certain key?	is it possible to write a song without a key?	1	4.04	-3.04
IV	*Antonyms*				
	chinese stocks close higher midday friday	chinese stocks open lower friday	1	3.9	-2.9
	the act of beginning something new.	the act of ending something.	0.8	3.67	-2.87
	higher than per cent but not very high.	lower than per cent but not very low.	1	4.22	-3.22
V	*Negation*				
	you are a christian.	therefore you are not a christian.	1.4	4.38	-2.98
	you should do it.	you should never do it.	1	4.56	-3.56
	it's not a good idea.	it's a good idea to do both.	1	3.9	-2.9

Table 1: Error analysis for MaxPool Spearman. Each entry contains a sentence pair, the gold similarity score, the scaled system similarity score, and the difference between the two scores. Errors are categorised into 5 types. The system heavily underestimates the similarity score when two sentences use different vocabulary yet have identical meaning (Type I). Inversely, it overestimates the similarity when the sentences use very related or even the same words but have different meaning (Types II & III). The similarity is also overestimated when two sentences contain antonyms, or when one sentence is a negation of the other (Types IV & V respectively).

		GloVe			fastText			word2vec		
		MPS	**DMX**	**Δ95% CI**	**MPS**	**DMX**	**Δ95% CI**	**MPS**	**DMX**	**Δ95% CI**
STS12	MSRpar	40.00	**49.41**	[-12.75, -6.26]	44.48	**48.94**	[-7.48, -1.53]	36.70	**41.74**	[-7.74, -2.40]
	MSRvid	**77.66**	71.92	[4.35, 7.27]	**82.44**	76.20	[5.00, 7.59]	74.34	**76.86**	[-3.65, -1.47]
	SMTeuroparl	**46.52**	**48.43**	[-4.61, 0.84]	50.18	**53.08**	[-5.44, -0.50]	**34.13**	28.03	[3.63, 8.38]
	surprise.OnWN	**69.23**	**69.86**	[-2.21, 0.93]	**73.12**	**72.79**	[-1.01, 1.70]	69.06	**71.26**	[-3.38, -1.04]
	surprise.SMTnews	**49.28**	**51.47**	[-5.37, 1.40]	**55.01**	**53.26**	[-1.72, 5.61]	45.09	**50.44**	[-8.09, -2.78]
STS13	FNWN	**46.16**	**39.79**	[-2.72, 15.43]	**44.14**	**42.34**	[-7.33, 11.19]	**49.66**	**42.34**	[-0.97, 17.12]
	headlines	**70.60**	**69.91**	[-0.75, 2.10]	**73.04**	**73.13**	[-1.26, 1.04]	**65.89**	**66.66**	[-1.97, 0.44]
	OnWN	**61.03**	52.12	[6.50, 11.66]	**71.37**	65.35	[3.97, 8.24]	**69.40**	**69.36**	[-1.23, 1.37]
STS14	deft-forum	**44.33**	**43.29**	[-2.29, 4.56]	**52.50**	47.16	[2.01, 8.80]	**45.60**	**47.27**	[-4.51, 1.22]
	deft-news	**70.69**	**70.55**	[-2.62, 2.82]	**70.64**	**71.04**	[-3.01, 2.01]	62.84	**65.84**	[-5.30, -0.82]
	headlines	**65.65**	**64.49**	[-0.44, 2.76]	**68.38**	**68.22**	[-1.09, 1.39]	62.00	**63.66**	[-3.03, -0.27]
	images	**78.98**	75.05	[2.44, 5.51]	**81.46**	79.39	[0.95, 3.24]	78.33	**80.51**	[-3.28, -1.14]
	OnWN	**69.20**	63.00	[4.46, 8.07]	**75.92**	72.83	[1.86, 4.44]	**75.35**	**75.43**	[-0.90, 0.80]
	tweet-news	**74.77**	**74.30**	[-1.49, 2.58]	76.38	**78.41**	[-3.39, -0.70]	71.17	**75.47**	[-5.70, -2.96]
STS15	answers-forums	**67.33**	61.94	[1.75, 9.45]	69.89	**73.57**	[-6.98, -0.42]	62.27	**66.44**	[-7.59, -0.83]
	answers-students	71.28	**73.53**	[-3.84, -0.71]	73.30	**75.82**	[-3.97, -1.12]	**74.20**	**75.07**	[-2.08, 0.25]
	belief	**72.83**	67.21	[2.28, 9.46]	**76.67**	**76.14**	[-2.21, 3.41]	**74.17**	**75.83**	[-3.67, 0.55]
	headlines	**72.14**	**72.26**	[-1.41, 1.08]	**74.78**	**74.45**	[-0.71, 1.31]	68.49	**69.95**	[-2.49, -0.46]
	images	**81.83**	79.30	[1.15, 3.98]	**84.84**	83.33	[0.50, 2.50]	82.60	**83.80**	[-2.08, -0.38]
STS16	answer-answer	**61.71**	**59.72**	[-1.64, 5.58]	**66.58**	**63.30**	[-1.24, 8.22]	**58.65**	**58.78**	[-3.09, 2.90]
	headlines	**70.57**	**71.71**	[-3.04, 0.80]	**72.81**	**73.40**	[-2.50, 1.10]	**67.87**	**68.18**	[-1.97, 1.38]
	plagiarism	**78.02**	**79.92**	[-4.41, 0.44]	**82.97**	**82.68**	[-1.61, 2.22]	**80.59**	**82.05**	[-3.11, 0.02]
	postediting	**81.11**	**80.48**	[-1.30, 2.97]	**82.31**	**84.15**	[-3.73, 0.15]	78.97	**81.73**	[-4.91, -1.13]
	question-question	**66.88**	**63.51**	[-0.58, 7.85]	**74.19**	69.71	[1.03, 8.16]	**66.79**	**65.74**	[-2.99, 4.85]

Table 2: **MaxPool Spearman vs. DynaMax:** Pearson correlations between human sentence similarity scores and a generated scores. Values in bold represent the best result for a subtask given a set of word vectors, based on a 95% BCa confidence interval (Efron, 1987) on the differences between the two correlations. In cases of no significant difference, both values are in bold.

		GloVe			fastText			word2vec		
		MPS	**ASP**	**Δ95% CI**	**MPS**	**ASP**	**Δ95% CI**	**MPS**	**ASP**	**Δ95% CI**
STS12	MSRpar	**40.00**	**35.90**	[-1.09, 9.00]	**44.48**	**39.66**	[-0.17, 9.82]	**36.70**	**38.79**	[-6.36, 2.07]
	MSRvid	**77.66**	68.80	[6.87, 11.04]	**82.44**	81.02	[0.22, 2.68]	74.34	**77.88**	[-4.98, -2.20]
	SMTeuroparl	**46.52**	**48.73**	[-5.82, 1.45]	**50.18**	**50.29**	[-3.85, 3.41]	**34.13**	16.96	[12.54, 21.09]
	surprise.OnWN	**69.23**	66.66	[0.12, 5.28]	**73.12**	**73.15**	[-1.79, 1.88]	**69.06**	**70.75**	[-3.64, 0.15]
	surprise.SMTnews	**49.28**	**47.12**	[-3.49, 7.65]	**55.01**	**56.67**	[-6.13, 3.14]	45.09	**53.93**	[-13.49, -4.14]
STS13	FNWN	**46.16**	**43.21**	[-8.69, 14.23]	**44.14**	**49.40**	[-14.85, 3.81]	**49.66**	**40.73**	[-0.61, 19.27]
	headlines	**70.60**	67.59	[1.16, 4.86]	**73.04**	**71.53**	[-0.12, 3.15]	**65.89**	**65.48**	[-1.29, 2.19]
	OnWN	**61.03**	57.66	[0.39, 6.44]	71.37	**74.33**	[-5.17, -1.00]	**69.40**	**67.49**	[-0.24, 4.08]
STS14	deft-forum	**44.33**	39.03	[0.24, 10.60]	**52.50**	46.20	[2.44, 10.69]	**45.60**	**42.95**	[-2.29, 9.03]
	deft-news	**70.69**	**68.99**	[-2.55, 5.86]	**70.64**	**73.08**	[-5.80, 0.72]	62.84	**67.33**	[-8.61, -0.31]
	headlines	**65.65**	61.87	[1.78, 5.86]	**68.38**	66.33	[0.37, 3.78]	**62.00**	**62.09**	[-1.86, 1.88]
	images	**78.98**	70.36	[6.31, 11.03]	**81.46**	**80.51**	[-0.52, 2.39]	**78.33**	**76.98**	[-0.42, 3.18]
	OnWN	**69.20**	**67.45**	[-0.09, 3.68]	75.92	**79.37**	[-4.78, -2.12]	**75.35**	**74.69**	[-0.76, 2.17]
	tweet-news	**74.77**	71.23	[0.93, 6.60]	**76.38**	**74.89**	[-0.51, 3.83]	**71.17**	**68.78**	[-0.32, 5.75]
STS15	answers-forums	**67.33**	50.25	[11.71, 22.87]	**69.89**	**68.28**	[-2.58, 5.53]	**62.27**	53.74	[3.31, 13.88]
	answers-students	**71.28**	**69.99**	[-1.16, 3.79]	**73.30**	**73.95**	[-2.72, 1.46]	**74.20**	**72.45**	[-0.12, 3.77]
	belief	**72.83**	58.77	[9.33, 20.04]	**76.67**	**73.71**	[-0.03, 6.03]	**74.17**	61.73	[8.05, 18.15]
	headlines	**72.14**	69.61	[0.86, 4.23]	**74.78**	72.93	[0.45, 3.18]	**68.49**	**68.58**	[-1.52, 1.40]
	images	**81.83**	73.85	[5.72, 10.37]	**84.84**	83.18	[0.33, 3.02]	**82.60**	80.04	[1.08, 4.07]
STS16	answer-answer	**61.71**	43.99	[10.25, 25.59]	**66.58**	54.51	[6.46, 18.13]	**58.65**	43.41	[8.65, 22.65]
	headlines	**70.57**	67.05	[0.74, 6.55]	**72.81**	**71.00**	[-1.11, 4.68]	**67.87**	**66.55**	[-1.12, 4.03]
	plagiarism	**78.02**	72.25	[1.47, 10.68]	**82.97**	**84.45**	[-4.11, 0.74]	**80.59**	75.21	[1.93, 9.23]
	postediting	**81.11**	69.03	[8.03, 17.65]	**82.31**	**82.73**	[-2.80, 2.06]	**78.97**	73.87	[1.53, 9.58]
	question-question	**66.88**	58.32	[1.26, 15.14]	**74.19**	**72.29**	[-1.78, 5.20]	**66.79**	**63.94**	[-3.32, 9.19]

Table 3: **MaxPool Spearman vs. MeanPool Spearman:** Pearson correlations between human sentence similarity scores and a generated scores. Values in bold represent the best result for a subtask given a set of word vectors, based on a 95% BCa confidence interval (Efron, 1987) on the differences between the two correlations. In cases of no significant difference, both values are in bold.

		GloVe			fastText			word2vec		
		MPS	**CKA**	**Δ95% CI**	**MPS**	**CKA**	**Δ95% CI**	**MPS**	**CKA**	**Δ95% CI**
STS12	MSRpar	**40.00**	**40.01**	[-2.97, 3.13]	**44.48**	**44.41**	[-2.53, 2.69]	**36.70**	**36.47**	[-2.42, 2.99]
	MSRvid	**77.66**	**76.81**	[-0.48, 2.28]	82.44	**84.45**	[-3.06, -1.01]	74.34	**79.95**	[-7.04, -4.33]
	SMTeuroparl	**46.52**	**48.94**	[-5.04, 0.24]	**50.18**	**51.36**	[-3.20, 0.78]	**34.13**	**35.28**	[-3.68, 1.15]
	surprise.OnWN	**69.23**	**67.86**	[-0.21, 2.89]	**73.12**	70.14	[1.44, 4.57]	**69.06**	**68.19**	[-0.54, 2.37]
	surprise.SMTnews	49.28	**53.80**	[-8.09, -1.03]	**55.01**	52.02	[0.03, 6.21]	45.09	**48.30**	[-6.28, -0.43]
STS13	FNWN	**46.16**	**36.36**	[-0.46, 20.49]	**44.14**	**43.61**	[-8.39, 9.81]	**49.66**	40.16	[0.38, 20.09]
	headlines	**70.60**	**71.85**	[-2.83, 0.28]	**73.04**	**73.61**	[-2.01, 0.87]	**65.89**	**64.66**	[-0.18, 2.77]
	OnWN	**61.03**	**60.95**	[-1.90, 2.17]	71.37	**74.25**	[-4.60, -1.38]	69.40	**72.06**	[-4.29, -1.20]
STS14	deft-forum	44.33	**50.65**	[-9.84, -2.84]	**52.50**	**54.16**	[-4.33, 1.15]	45.60	**52.17**	[-9.61, -3.73]
	deft-news	**70.69**	**73.44**	[-5.59, 0.40]	**70.64**	**73.06**	[-5.57, 0.50]	62.84	**67.26**	[-7.53, -1.56]
	headlines	**65.65**	**66.32**	[-2.24, 0.89]	**68.38**	**68.45**	[-1.61, 1.51]	**62.00**	**61.54**	[-0.91, 1.93]
	images	**78.98**	77.47	[0.08, 2.95]	**81.46**	**81.76**	[-1.42, 0.76]	78.33	**80.57**	[-3.49, -1.07]
	OnWN	**69.20**	**69.16**	[-1.44, 1.52]	75.92	**78.46**	[-3.80, -1.36]	75.35	**77.00**	[-2.72, -0.61]
	tweet-news	**74.77**	**73.95**	[-0.92, 2.84]	**76.38**	73.41	[1.26, 4.81]	**71.17**	**71.86**	[-2.28, 1.03]
STS15	answers-forums	**67.33**	**66.48**	[-2.52, 4.20]	69.89	**72.78**	[-5.92, -0.14]	**62.27**	**64.01**	[-5.00, 1.36]
	answers-students	**71.28**	**72.75**	[-3.22, 0.22]	**73.30**	**71.92**	[-0.23, 3.11]	**74.20**	**73.59**	[-0.86, 2.25]
	belief	**72.83**	**71.56**	[-2.26, 5.44]	**76.67**	**76.00**	[-1.28, 2.84]	**74.17**	**74.16**	[-2.56, 3.22]
	headlines	72.14	**74.05**	[-3.12, -0.71]	**74.78**	**75.58**	[-2.01, 0.35]	**68.49**	**69.00**	[-1.66, 0.68]
	images	**81.83**	**81.35**	[-0.97, 2.00]	**84.84**	**85.40**	[-1.52, 0.38]	82.60	**84.02**	[-2.37, -0.50]
STS16	answer-answer	**61.71**	56.04	[0.91, 10.55]	**66.58**	61.81	[1.28, 8.66]	**58.65**	51.21	[2.99, 12.68]
	headlines	**70.57**	**70.83**	[-2.46, 1.94]	**72.81**	**72.87**	[-2.50, 2.12]	**67.87**	65.52	[0.39, 4.59]
	plagiarism	**78.02**	**79.36**	[-3.91, 1.41]	**82.97**	79.83	[0.66, 5.92]	**80.59**	**80.66**	[-1.99, 1.86]
	postediting	**81.11**	**79.94**	[-1.25, 4.32]	**82.31**	80.40	[0.16, 3.85]	**78.97**	**78.86**	[-2.01, 2.14]
	question-question	66.88	**72.01**	[-9.29, -1.58]	**74.19**	**73.68**	[-2.15, 3.23]	66.79	**70.59**	[-8.03, -0.16]

Table 4: **MaxPool Spearman vs. CKA Gaussian:** Pearson correlations between human sentence similarity scores and a generated scores. Generated scores Values in bold represent the best result for a subtask given a set of word vectors, based on a 95% BCa confidence interval (Efron, 1987) on the differences between the two correlations. In cases of no significant difference, both values are in bold.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393. Association for Computational Linguistics.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 982–991, Honolulu, Hawaii. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. 2019. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. In *International Conference on Learning Representations*.