ACL 2022

Just Rank: Rethinking Evaluation with Word and Sentence Similarities

Bin Wang, C.-C. Jay Kuo, Haizhou Li

NUS National University of Singapore

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

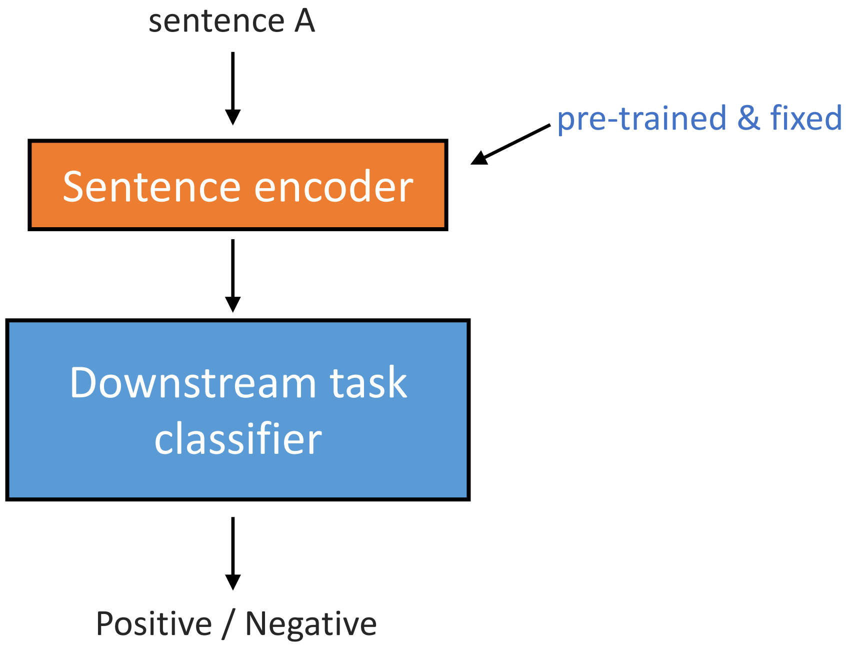
USC

Kriston AI

Embedding Evaluation (word & sentence)

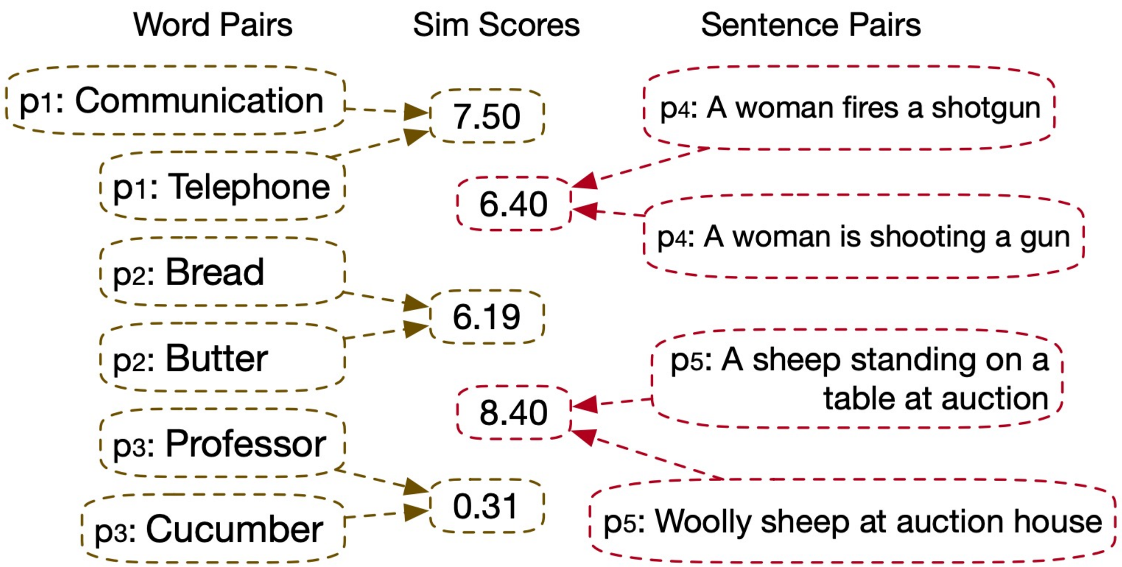
- Intrinsic evaluation
  - Word and sentence similarity (most popular)
  - Analogy tasks
  - Probing tasks

- Downstream tasks
  - Sentiment/topic classification
  - Natural language inference



Similarity Evaluation Scheme

- Human-labeled pair similarities
- Embedding for samples
- Does embedding similarities match human similarities?



Word Pairs

Sim Scores

Sentence Pairs

p1: Communication

p1: Telephone

7.50

p2: Bread

p2: Butter

6.19

p3: Professor

p3: Cucumber

0.31

p4: A woman fires a shotgun

p4: A woman is shooting a gun

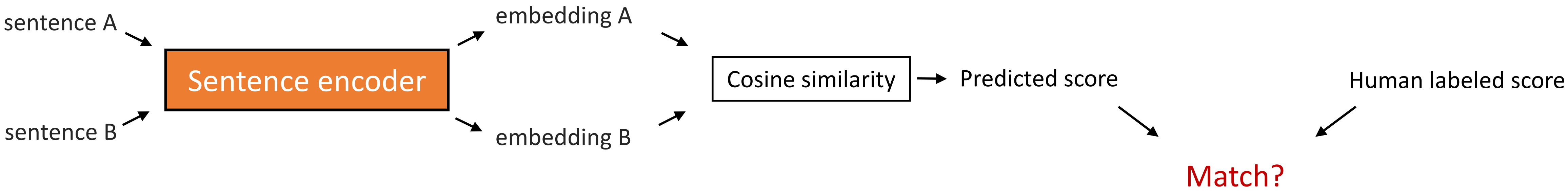
6.40

p5: A sheep standing on a table at auction

p5: Woolly sheep at auction house

8.40

<u>Pipeline</u>:



sentence A

sentence B

Sentence encoder

embedding A

embedding B

Cosine similarity

Predicted score

Human labeled score

Match?

Outline

- Problems with current similarity evaluation
  - Multifaceted relations
  - Weak correlation with downstream tasks
  - Overfitting to similarity metrics and whitening tricks

- A new evaluation paradigm – *EvalRank*
  - Spreading-Activation Theory (SAT)
  - Dataset and methodology
  - Experimental results

Multifaceted relations

- Concept of similarity and relatedness is not well defined

Similarity-level:

synonym > hypernym > antonym

Relatedness-level:

synonym > hypernym ≈ antonym

- Annotation process is not intuitive to humans
  - Instructions are not clear
  - Human perceptions are not unique
  - Alternatives: priming stimulus, comparative annotations[1]

Pairs with score 2:

"share some details"

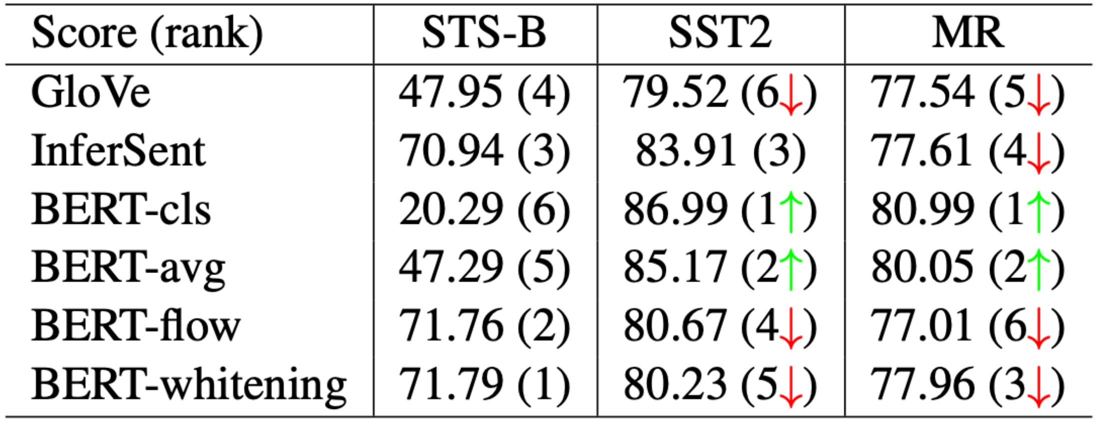
Pairs with score 1:

"on the same topic"

[1] Abdalla, Mohamed, Krishnapriya Vishnubhotla, and Saif M. Mohammad. "What Makes Sentences Semantically Related: A Textual Relatedness Dataset and Empirical Study." arXiv preprint arXiv:2110.04845 (2021).

Weak corr. w/ downstream

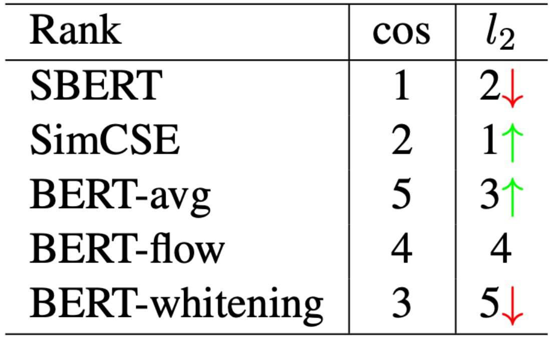
- Good performance on similarity tasks does not guarantee good performance on downstream tasks
  - Different properties of interest
    - Mimic human perception *V.S.* Real-world application
  - Different ways of inference
    - Simple metric (cosine, l2) *V.S.* Non-linear classifier (MLP, LSTM, Transformers)



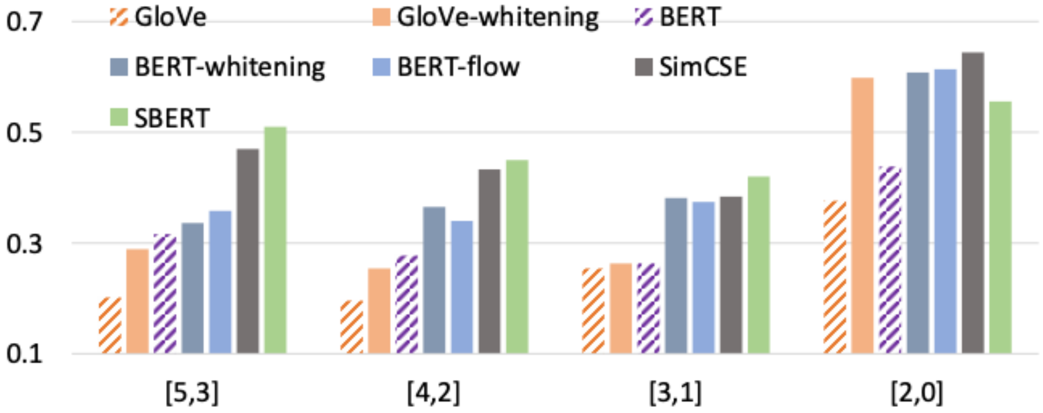
Score (rank)	STS-B	SST2	MR
GloVe	47.95 (4)	79.52 (6↓)	77.54 (5↓)
InferSent	70.94 (3)	83.91 (3)	77.61 (4↓)
BERT-cls	20.29 (6)	86.99 (1↑)	80.99 (1↑)
BERT-avg	47.29 (5)	85.17 (2↑)	80.05 (2↑)
BERT-flow	71.76 (2)	80.67 (4↓)	77.01 (6↓)
BERT-whitening	71.79 (1)	80.23 (5↓)	77.96 (3↓)

Overfitting

- Current models are optimizing towards certain evaluation metrics
  - Cosine similarity
  - Whitening tricks
    - Helps with similarity with cosine
    - Little/no help for similarity with l2 distance
    - Little/no help for downstream tasks

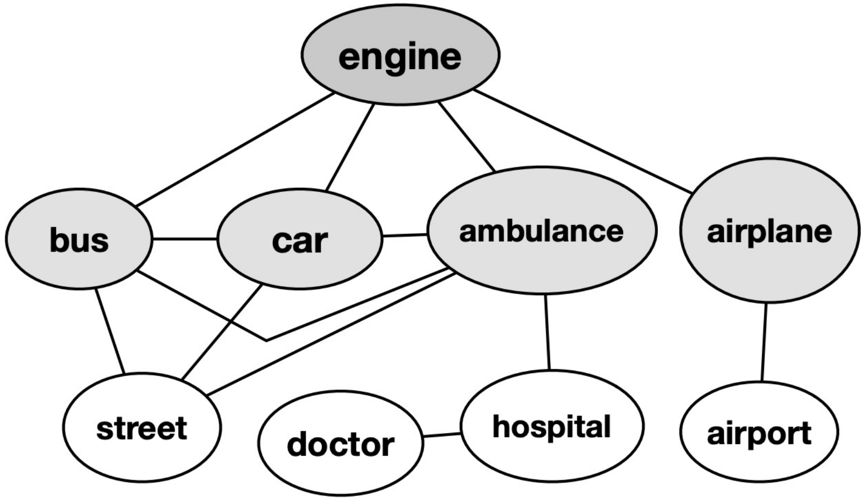


Rank	cos	l_2
SBERT	1	2↓
SimCSE	2	1↑
BERT-avg	5	3↑
BERT-flow	4	4
BERT-whitening	3	5↓



EvalRank - Motivation

- Concept network in Spread Activation Theory (SAT)[2]
  - Most similar pairs are less noisy
  - Measurable by simple distance metrics (cosine, l2)
  - More important to downstream tasks

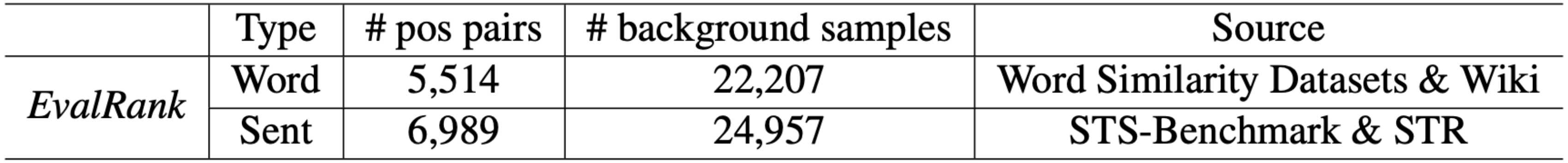


An example of Concept Network in SAT

[2] Collins, Allan M., and Elizabeth F. Loftus. "A spreading-activation theory of semantic processing." Psychological review 82.6 (1975): 407.

EvalRank - Methodology

- Dataset
  - 25% from word & sentence similarity datasets



	Type	# pos pairs	# background samples	Source
EvalRank	Word	5,514	22,207	Word Similarity Datasets & Wiki
	Sent	6,989	24,957	STS-Benchmark & STR

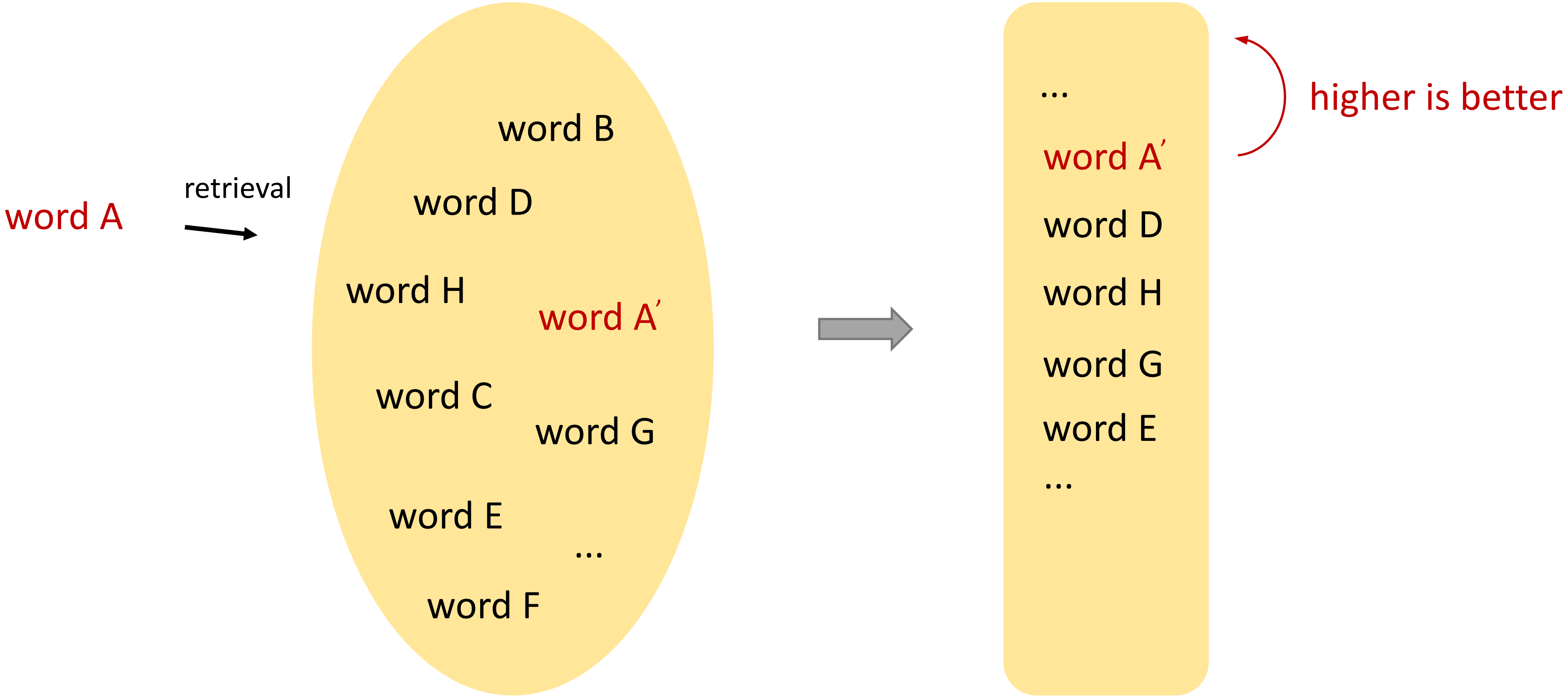
- Retrieval-based ranking

$$rank_i = rank(S(c_x, c_y), [||_{j=1, j \neq x}^{n} S(c_x, c_j)])$$

$$MRR = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{rank_i}$$

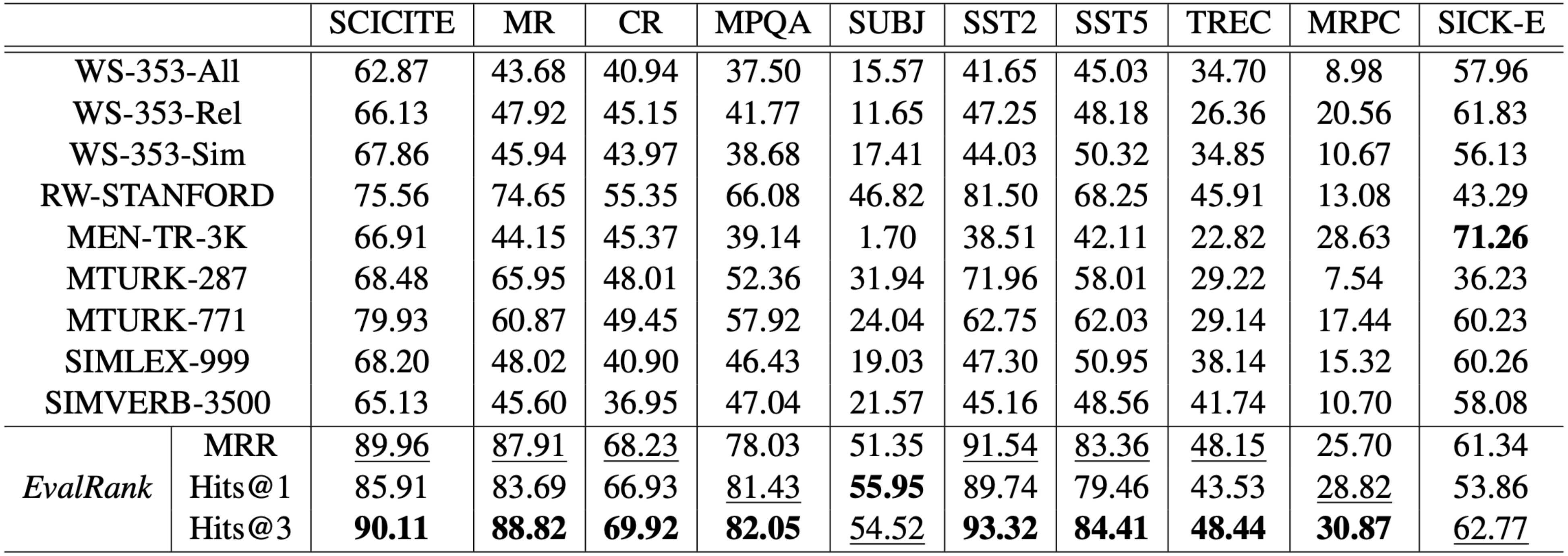
$$Hits@k = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[rank_i \leq k]$$

EvalRank - Methodology



EvalRank – Experimental Results

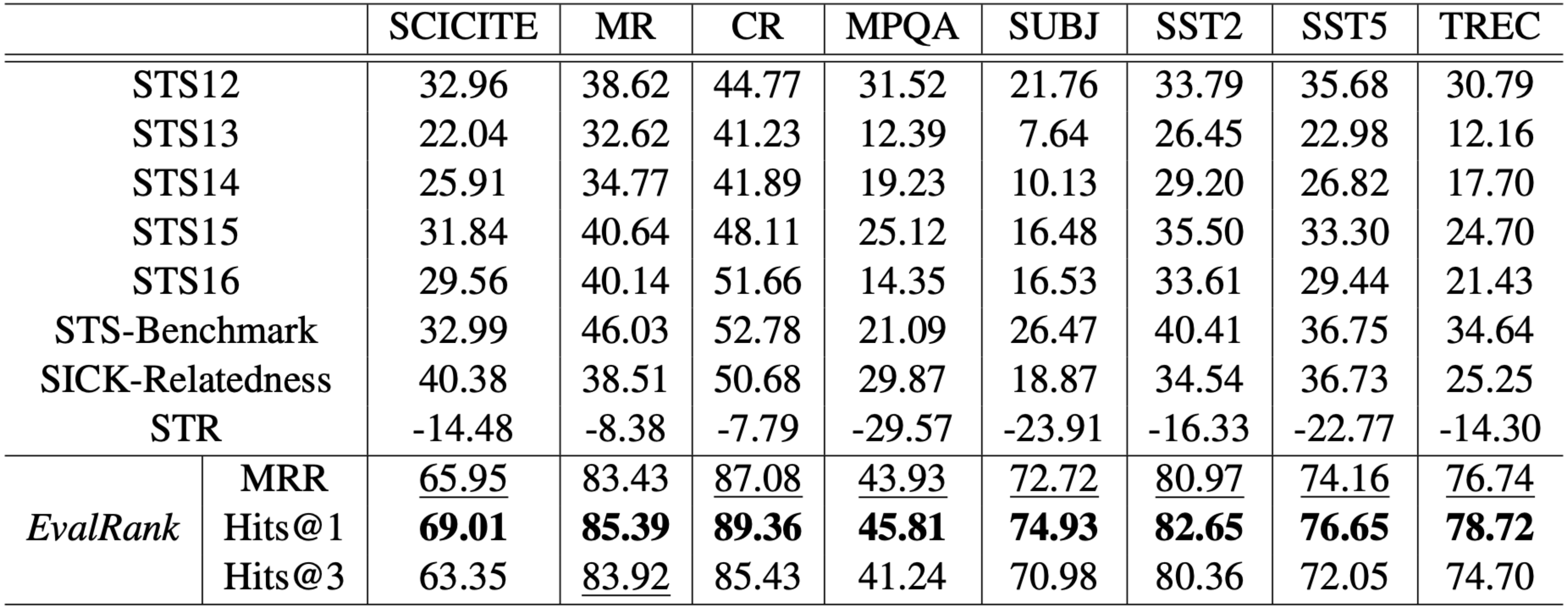
- Word-level (38 word embedding variants)



		SCICITE	MR	CR	MPQA	SUBJ	SST2	SST5	TREC	MRPC	SICK-E
WS-353-All		62.87	43.68	40.94	37.50	15.57	41.65	45.03	34.70	8.98	57.96
WS-353-Rel		66.13	47.92	45.15	41.77	11.65	47.25	48.18	26.36	20.56	61.83
WS-353-Sim		67.86	45.94	43.97	38.68	17.41	44.03	50.32	34.85	10.67	56.13
RW-STANFORD		75.56	74.65	55.35	66.08	46.82	81.50	68.25	45.91	13.08	43.29
MEN-TR-3K		66.91	44.15	45.37	39.14	1.70	38.51	42.11	22.82	28.63	**71.26**
MTURK-287		68.48	65.95	48.01	52.36	31.94	71.96	58.01	29.22	7.54	36.23
MTURK-771		79.93	60.87	49.45	57.92	24.04	62.75	62.03	29.14	17.44	60.23
SIMLEX-999		68.20	48.02	40.90	46.43	19.03	47.30	50.95	38.14	15.32	60.26
SIMVERB-3500		65.13	45.60	36.95	47.04	21.57	45.16	48.56	41.74	10.70	58.08
EvalRank	MRR	89.96	87.91	68.23	78.03	51.35	91.54	83.36	48.15	25.70	61.34
	Hits@1	85.91	83.69	66.93	81.43	**55.95**	89.74	79.46	43.53	28.82	53.86
	Hits@3	**90.11**	**88.82**	**69.92**	**82.05**	54.52	**93.32**	**84.41**	**48.44**	**30.87**	62.77

EvalRank – Experimental Results

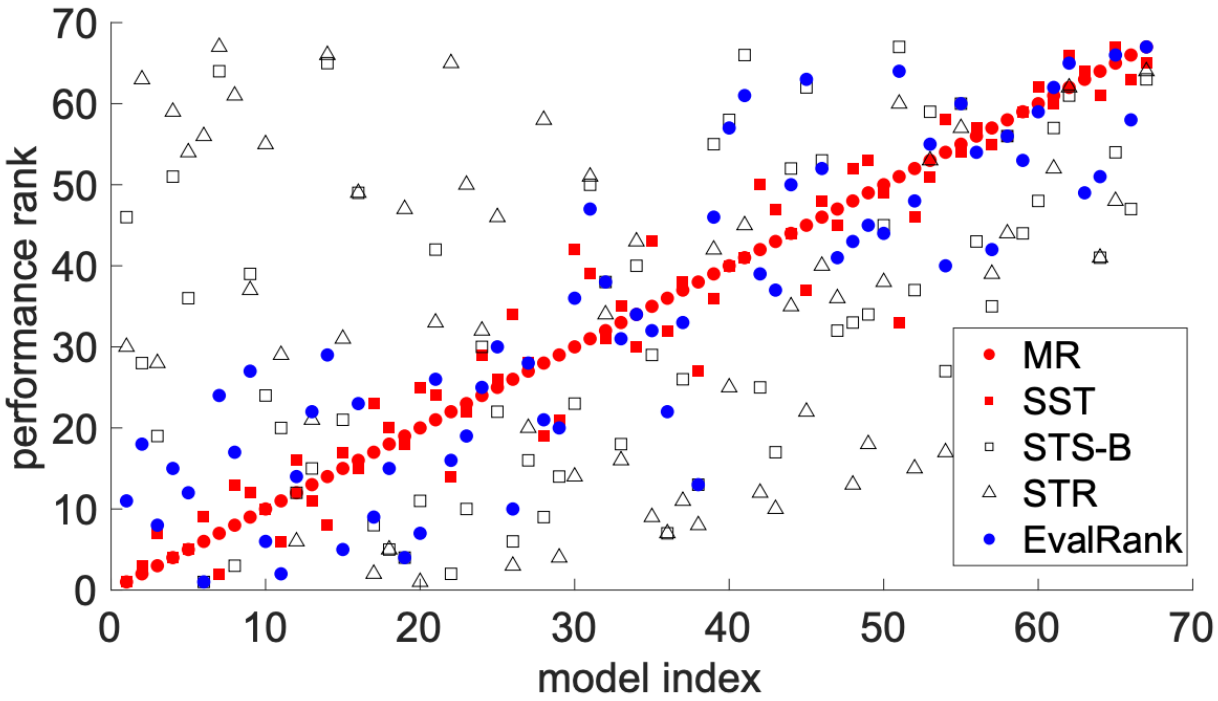
- Sentence-level (67 sentence embedding variants)



		SCICITE	MR	CR	MPQA	SUBJ	SST2	SST5	TREC
STS12		32.96	38.62	44.77	31.52	21.76	33.79	35.68	30.79
STS13		22.04	32.62	41.23	12.39	7.64	26.45	22.98	12.16
STS14		25.91	34.77	41.89	19.23	10.13	29.20	26.82	17.70
STS15		31.84	40.64	48.11	25.12	16.48	35.50	33.30	24.70
STS16		29.56	40.14	51.66	14.35	16.53	33.61	29.44	21.43
STS-Benchmark		32.99	46.03	52.78	21.09	26.47	40.41	36.75	34.64
SICK-Relatedness		40.38	38.51	50.68	29.87	18.87	34.54	36.73	25.25
STR		-14.48	-8.38	-7.79	-29.57	-23.91	-16.33	-22.77	-14.30
EvalRank	MRR	65.95	83.43	87.08	43.93	72.72	80.97	74.16	76.74
	Hits@1	**69.01**	**85.39**	**89.36**	**45.81**	**74.93**	**82.65**	**76.65**	**78.72**
	Hits@3	63.35	83.92	85.43	41.24	70.98	80.36	72.05	74.70

EvalRank – Experimental Results

- Visualization



  - EvalRank correlation better with MR & SST

Take-home messages and future work

- Possible problems with similarity evaluation
  - Mimic human perception
  - Fixed evaluation paradigm
  - Focus on single intrinsic evaluation may hinder the improvement of embedding models
- New intrinsic evaluation – *EvalRank*
  - Better correlation with downstream tasks
- Future work
  - New intrinsic datasets
  - Multifaceted embeddings

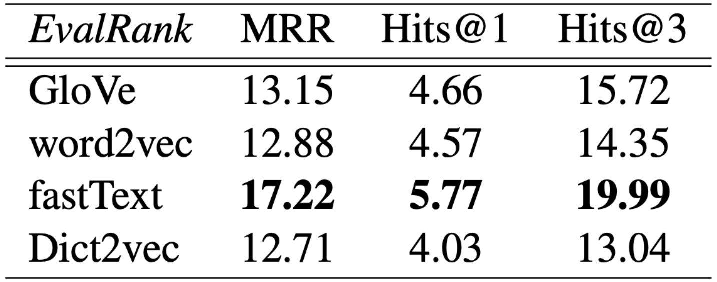
Thank you!

- Evaluation toolkit publicly available

- Support a series of embedding architectures

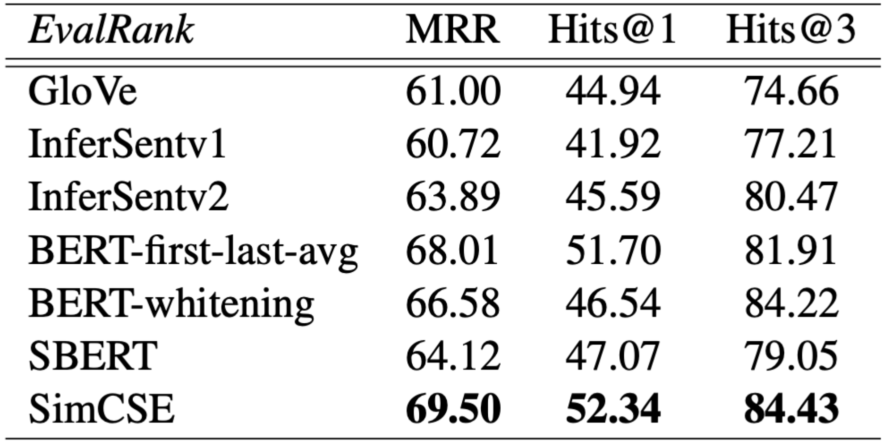
- Benchmarking results

https://github.com/BinWang28/EvalRank-Embedding-Evaluation



EvalRank	MRR	Hits@1	Hits@3
GloVe	13.15	4.66	15.72
word2vec	12.88	4.57	14.35
fastText	**17.22**	**5.77**	**19.99**
Dict2vec	12.71	4.03	13.04

Word-level benchmarking



EvalRank	MRR	Hits@1	Hits@3
GloVe	61.00	44.94	74.66
InferSentv1	60.72	41.92	77.21
InferSentv2	63.89	45.59	80.47
BERT-first-last-avg	68.01	51.70	81.91
BERT-whitening	66.58	46.54	84.22
SBERT	64.12	47.07	79.05
SimCSE	**69.50**	**52.34**	**84.43**

Sentence-level benchmarking