Investigating the Effects of Word Substitution Errors on Sentence Embeddings

Rohit Voleti[1], Julie M. Liss[2], Visar Berisha[1,2]

[1]School of Electrical, Computer, & Energy Engineering, ASU
[2]Department of Speech & Hearing Science, ASU

1. Introduction:

- Word and sentence *embeddings* (vector representations) are usually trained & evaluated on corpora with perfect text transcriptions
- Applications often rely on automatic speech recognition (ASR)
- We propose a simple word substitution error simulator
  - **Goal**: Realistically corrupt clean text to evaluate models on noisy data with a given *word error rate* (WER)
- We evaluate the performance of several *sentence embeddings* after introducing substitution errors on *semantic textual similarity* (STS)

2. ASR Word Substitution Error Simulator:

- ASR word determinations typically rely on phonemic similarity and a language model which determines semantically plausible confusions
- Our simulator:
  - Models phonemics using *phonological edit distance* [1, 2, 3]
  - Models semantics with *GloVe* [4] word embeddings
  - Define $d_{ij} = f(d_{ij}^S, d_{ij}^P)$ as distance, i.e. *semantic* & *phonemic*

$$d_{ij}^S = 1 - \cos\theta_{ij} = 1 - \frac{\mathbf{w}_i^T \mathbf{w}_j}{\|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2} \quad \text{(\textit{GloVe} cosine dist.)}$$

$$d_{ij}^P = \text{PhonEdtDist}(\mathbf{w}_i, \mathbf{w}_j) \quad \text{(ARPABET transcriptions)}$$

Algorithm 1 Random replacement of words in a given a corpus with a specified WER to simulate realistic ASR errors.

procedure CORRUPT SENTENCES(corpus, WER)
  Find all unique tokens, w_i, in the corpus that exist in the set of pre-trained *GloVe* embeddings
  Filter all w_i to those in pronouncing dictionary
  for each w_i **do**
    Find w_j, $j = 1, \cdots, N$ most similar words by d_{ij}^S
    ARPABET transcription for w_i, all w_j
    for each w_j **do**
      Compute d_{ij}^P from w_i to w_j, where $j = 1, \cdots, N$
    end for
    Keep only M values of $d_{ij}^P \leq$ thresh, where $M \leq N$
    for $j = 1, \cdots, M$ **do**
      Compute $P_{\text{subs}}(w_j|w_i) = \alpha \cdot \exp(-d_{ij}/\sigma^2)$
    end for
  end for
  Randomly select words to replace given WER
  Replace selected words with error words based on the 0.5cm probability distributions computed
end procedure

3. Sentence Embeddings Evaluated:

- Unweighted average of word2vec (w2v) [5] vectors
  - With and without *stop words* removed
- Smooth Inverse Frequency (SIF) [6]
- Unsupervised Smooth Inverse Frequency (uSIF) [7]
- Low-rank Subspace [8]
- InferSent, based on FastText [9]

We evaluated the performance of several **sentence embedding** models after *simulating* ASR-plausible **substitution errors** on perfectly transcribed text.

Original Sentence	Corrupted Sentence
Obama holds out over Syria strikes	Obama *helps* out *every* *Sharia* strikes
Russia warns Ukraine against EU deal	Russia warns *Euro* against EU deal
Gov. Linda Lingle and members of her staff were at the Navy base and watched the launch.	Gov . *Cindy* Lingle *add* *mentors* of her *staffs* were at the *NASA* base *add* watched the *launcher* .
I have had the same problem.	*Eyes* have had the same *progress*.
A white cat looking out of a window.	A white cat *letting* out of a window.

Key Findings for *Sentence Embeddings* and *Semantic Textual Similarity*:

- Unweighted averaging of *word2vec* vectors is ***least*** impacted by introducing errors
  - However, STS performance is ***most negatively impacted***
- *Smooth Inverse Frequency* (SIF) / *Unsupervised SIF* are ***most*** impacted by errors
  - ***More robust*** on STS performance than unweighted averages
- *InferSent* is moderately affected by introduced errors
  - ***Most robust*** on STS performance, least impacted by errors
  - More complicated to learn (deep LSTM architecture) but best at handling errors

References:

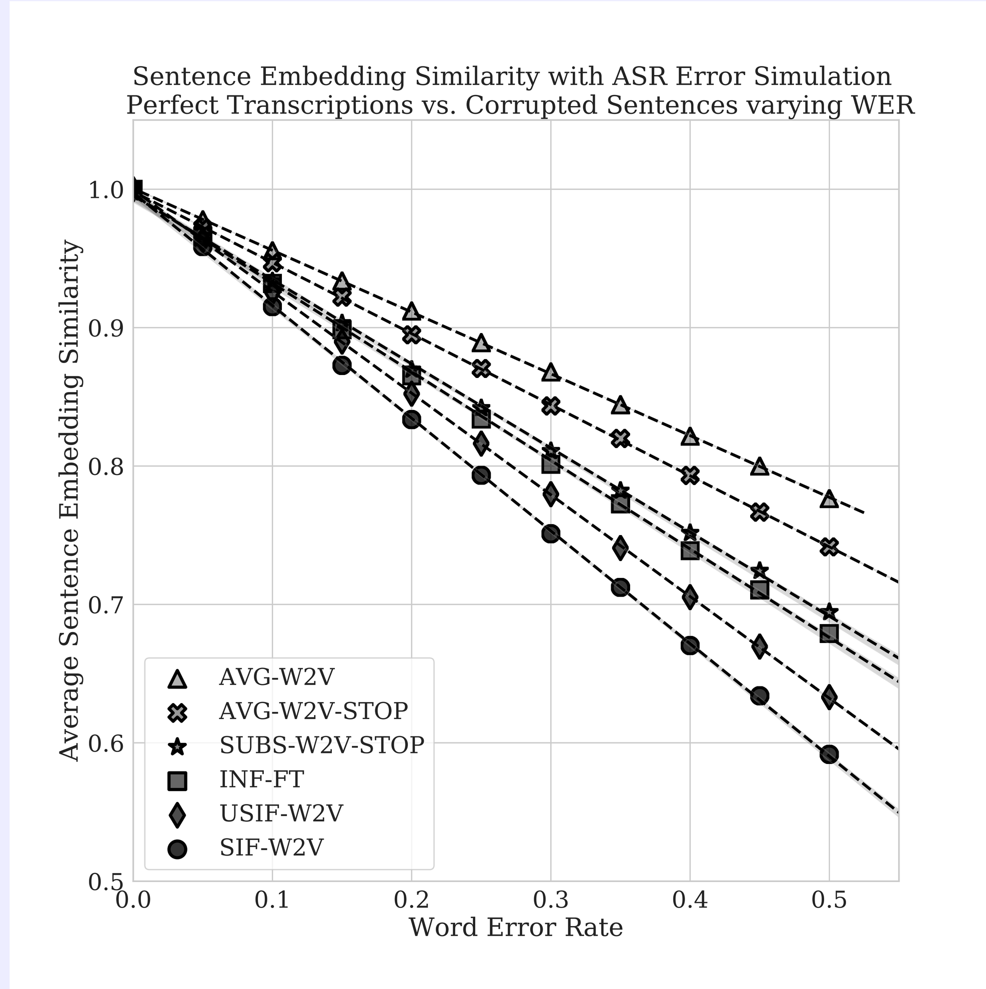
[1] Nathan C. Sanders and Steven B. Chin, "Phonological Distance Measures*," *Journal of Quantitative Linguistics*, vol. 16, no. 1, pp. 96–114, Feb. 2009.
[2] Blake Allen and Michael Becker, "Learning alternations from surface forms with sublexical phonology," *Unpublished manuscript, University of British Columbia and Stony Brook University. Available as lingbuzz/002503*, 2015.
[3] Kathleen Currie Hall,Blake Allen,Michael Fry,Scott Mackie, and Michael McAuliffe, "Phonological CorpusTools," in *14th Conference for Laboratory Phonology*, Tokyo, Japan, 2015.
[4] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global Vectors for Word Representation," 2014, pp. 1532–1543, Association for Computational Linguistics.
[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[6] Sanjeev Arora, Yingyu Liang, and Tengyu Ma, "A Simple but Tough-to-Beat Baseline for Sentence Embeddings," in *Proceedings of 5th International Conference on Learning Representations*, Toulon, France, 2017, p. 16.
[7] Kawin Ethayarajh, "Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline," in *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018, pp. 91–100.
[8] Jiaqi Mu, Suma Bhat, and Pramod Viswanath, "Representing Sentences as Low-Rank Subspaces," *arXiv:1704.05358 [cs]*, Apr. 2017
[9] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *arXiv:1705.02364 [cs]*, May 2017.
[10] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," 2017, pp. 1–14, Association for Computational Linguistics.
[11] M Marelli, S Menini, M Baroni, L Bentivogli, R Bernardi, and R Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," p. 8

SCAN for Full Paper on IEEE Xplore

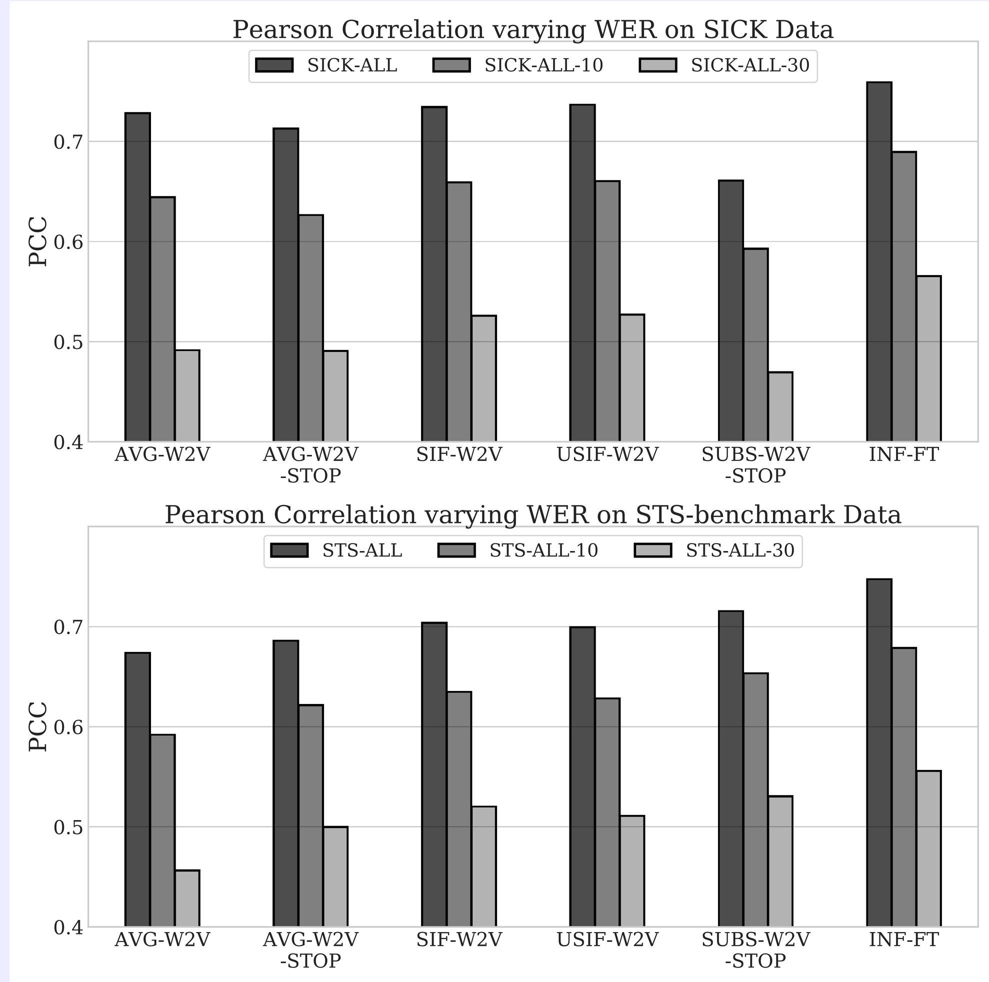
4. Results:

Using sentence pairs from *STS-benchmark* [10] and *SICK* [11]

A. Comparing the similarity of clean sentences to their corrupted versions, varying Word Error Rate (WER) from 0% to 50%



B. Semantic similarity of sentence pairs (STS) from both datasets
- 3 different WER (0%, 10%, 30%)
- *Pearson Corr. Coeff.* (PCC) between computed & human scores



Sentence Embedding	STS Corpus (dev & test set)	$\mathbf{PCC_{0\%}}$ / $\mathbf{PCC_{10\%}}$ / $\mathbf{PCC_{30\%}}$ (×100)	$\mathbf{PCC_{30\%}/PCC_{0\%}}$
AVG-W2V:	SICK:	72.84 / 64.44 / 49.18	67.52%
	STS-benchmark:	67.40 / 59.23 / 45.64	67.72%
AVG-W2V-STOP:	SICK:	71.30 / 62.67 / 49.09	68.85%
	STS-benchmark:	68.61 / 62.15 / 49.99	72.85%
SIF-W2V:	SICK:	73.44 / 65.93 / 52.60	71.63%
	STS-benchmark:	70.39 / 63.51 / 52.06	73.96%
USIF-W2V:	SICK:	73.70 / 66.06 / 52.71	71.51%
	STS-benchmark:	69.95/ 62.85 / 51.11	73.07%
SUBS-W2V-STOP:	SICK:	66.10 / 59.28 / 46.94	71.02%
	STS-benchmark:	71.58 / 65.36 / 53.05	74.10%
INF-FT:	SICK:	75.94 / 68.95 / 56.56	74.48%
	STS-benchmark:	74.77 / 67.88 / 55.60	74.36%

Arizona State University