

INVESTIGATING THE EFFECTS OF WORD SUBSTITUTION ERRORS ON SENTENCE EMBEDDINGS

Rohit Voleti¹, Julie M. Liss², Visar Berisha^{1,2}

Arizona State University
Department of Electrical, Computer, & Energy Engineering¹
Department of Speech & Hearing Science²
Tempe, AZ, USA

ABSTRACT

A key initial step in several natural language processing (NLP) tasks involves embedding phrases of text to vectors of real numbers that preserve semantic meaning. To that end, several methods have been recently proposed with impressive results on semantic similarity tasks. However, all of these approaches assume that perfect transcripts are available when generating the embeddings. While this is a reasonable assumption for analysis of written text, it is limiting for analysis of transcribed text. In this paper we investigate the effects of word substitution errors, such as those coming from automatic speech recognition errors (ASR), on several state-of-the-art sentence embedding methods. To do this, we propose a new simulator that allows the experimenter to induce ASR-plausible word substitution errors in a corpus at a desired word error rate. We use this simulator to evaluate the robustness of several sentence embedding methods. Our results show that pre-trained neural sentence encoders are both robust to ASR errors and perform well on textual similarity tasks after errors are introduced. Meanwhile, unweighted averages of word vectors perform well with perfect transcripts, but their performance degrades rapidly on textual similarity tasks for text with word substitution errors.

Index Terms— Sentence Embeddings, Speech Recognition, Natural Language Processing, Semantic Embedding, ASR Error Simulator

1. INTRODUCTION & RELATED WORK

Many real-world applications motivate the need to accurately capture the semantic content of a sentence. Examples include sentiment analysis of product reviews, customer service chatbots, biomedical informatics, among several others. *Word embeddings* map words from a lexicon to a continuous vector space in which nearby vectors are also semantically related. Similarly, *sentence embeddings* map individual phrases or sentences to a continuous vector space that preserve the text semantics. The approaches to the word-embedding problem range from simple singular value decomposition of co-occurrence matrices [1] to neural network models trained on large corpora (e.g. *word2vec* [2], *GloVe* [3], and *FastText* [4]).

These approaches have revolutionized NLP research by showing impressive results on downstream NLP tasks; however, to the best of our knowledge, all of the previous work on sentence and word embeddings is built upon the assumption that the available text for training and testing each embedding model is perfectly transcribed. In most real-world applications, it is unlikely that textual language data will be free of error. In fact, an increasing number of applica-

tions rely on *automatic speech recognition* (ASR) systems for transcriptions. The performance of an ASR system can be characterized by its *word-error rate* (WER), which defines the percentage of incorrect word errors given by the output of a particular system. Typical modern ASR systems have a WER ranging from $\sim 10\%$ to $\sim 35\%$ [5]. With a few exceptions, i.e. [6], [7], [8], [9], the effects of ASR errors have been largely ignored in many NLP applications. And, to the best of our knowledge, no previous work has been conducted to evaluate the effects of ASR errors on sentence embeddings and their performance in downstream NLP tasks.

In this work, we evaluate the robustness of several state-of-the-art sentence embeddings to word substitution errors typical of ASR systems¹. To do this, we propose a new method for simulating realistic ASR transcription errors with a specified WER that is implemented with only publicly available tools for acoustic and semantic modeling. We evaluate the resultant embeddings on the semantic textual similarity (STS) task, a popular research topic in NLP within the area of statistical distributional semantics. In STS, the goal is to develop sentence embeddings that can successfully model the semantic similarity between two sentences (or another arbitrary collection of words). Several recently developed sentence embedding methods have shown very promising results on STS tasks [2], [3], [10], [11], [12], [13], [14], [15]; however, all have been evaluated using perfect transcripts. We attempt to re-evaluate the results on standard STS datasets after introducing the errors simulated using our approach. In short, the contributions of this work are: 1) a new simulator for introducing ASR-plausible word substitution errors that utilizes phonetic and semantic information to randomly replace words in a corpus with likely confusion words, 2) an evaluation of five recent sentence embedding methods and their robustness to simulated ASR noise, and 3) an evaluation of the STS performance of these sentence embeddings with simulated ASR errors and a variable WER using the *SICK* [16] and *STS-benchmark* [17] datasets.

2. WORD SUBSTITUTION ERROR SIMULATION

In this section we propose a new word substitution error simulator intended to model plausible substitutions that an ASR algorithm might produce. Our approach is based on the observation that the nature of word substitution errors in ASR systems depends on the phonemic distance between the true word and the substituted word (because of the underlying acoustic model) and on the semantic distance be-

¹WER calculation includes unintended word *insertions*, *deletions*, and *substitutions*. We note that a limitation of our model is that it only considers potential substitution errors when simulating ASR error

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

Table 1. Example sentence pairs from STS-benchmark [17] and SICK corpora [18] after corrupting all sentences with WER of 30%. Substituted word errors are shown in italics. A high WER is used here to demonstrate the types of substitution errors simulated by our method, incorporating both semantic and phonemic distance measures.

tween the true word and the substituted word (because of the underlying language model). To that end, we define the probability of substituting word w_i with word w_j by

$$P_{\text{subs}}(w_j|w_i) = \alpha \cdot \exp\left(-\frac{d_{ij}}{\sigma^2}\right), \quad (1)$$

where d_{ij} is a notion of distance between w_i and w_j comprised of both the phonemic and semantic distance, σ is a user-defined parameter that controls the shape of the resulting probability mass function (PMF), and α is a normalization constant that makes the marginal PMF in Equation 1 sum to one for each given w_i .

Estimating the substitution probabilities: Given a corpus for which we want to simulate word substitution errors, we first compute the set of all unique words. Next, we consider the pair-wise substitution error probabilities using Eqn. (1). Estimating the probability of a substitution requires that we estimate d_{ij} . Loosely speaking, we model the total distance as being comprised of a phonemic distance between the words (contribution of acoustic model in ASR) and a semantic distance between words (contribution of the language model in ASR).

To estimate the phonemic distance, we use a phonological edit distance between words w_i and w_j , d_{ij}^P [19], [20], [21], loosely based on the Levenshtein edit distance [22], which compares the number of single-character edits one string would need to be identical to another string. We consider ARPABET transcriptions based on the *CMU Pronouncing Dictionary* [23] to similarly compute phonemic similarity. To encode each phoneme, we use the *articulation features* provided by Hayes in [24]. The result is a binary feature matrix for each English phoneme in ARPABET. The phonological edit distance between two words can be computed as the number of *single-feature* edits that are required to pronounce the first word like the second, as outlined by Sanders *et al.* in [19].

To estimate the semantic distance between the words, we use the *GloVe* embeddings [3] for every word in the corpus and estimate the pairwise *cosine distance* as

$$d_{ij}^S = 1 - \cos \theta_{ij} = 1 - \frac{\mathbf{w}_1^T \mathbf{w}_2}{\|\mathbf{w}_1\|_2 \|\mathbf{w}_2\|_2}, \quad (2)$$

where \mathbf{w}_i and \mathbf{w}_j represent the vector representations of two distinct words w_i and w_j , and θ_{ij} represents the angle between the vectors.

Algorithm implementation: The total distance in Equation 1 can be modeled using some function of the two contributions discussed above, $d_{ij} = f(d_{ij}^S, d_{ij}^P)$. However, this approach requires that we estimate the conditional probability in Equation 1 for every pair of words in a corpus; for large, realistic vocabulary sizes, this becomes prohibitively large.

To alleviate the need to estimate all pairwise probabilities, we only consider the $N = 1000$ semantically most similar words in the corpus using d_{ij}^S and estimate the marginal distribution for that subset of words, assuming that it is zero for all others. In addition, in Equation 1, we model d_{ij} using only the contribution from the phonological edit distance. The parameter σ can be chosen and tuned based on empirical results. We found that setting σ equal to the average phonological edit distance between each cluster of potential replacement words and the target word provided reasonable results. The overall procedure is summarized in Algorithm 1.

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

```

1: procedure CORRUPT SENTENCES(corpus, WER)
2:   Find all unique tokens,  $w_i$ , in the corpus that exist in the set
     of pre-trained GloVe embeddings
3:   Filter all  $w_i$  to those in pronouncing dictionary
4:   for each  $w_i$  do
5:     Find  $w_j$ ,  $j = 1, \dots, N$  most similar words by  $d_{ij}^S$ 
6:     ARPABET transcription for  $w_i$ , all  $w_j$   $\triangleright$  CMU Dict
7:     for each  $w_j$  do
8:       Compute  $d_{ij}^P$  from  $w_i$  to  $w_j$ , where  $j = 1, \dots, N$ 
9:       Keep only  $M$  values of  $d_{ij}^P \leq \text{thresh}$ , where  $M \leq N$ 
10:      for  $j = 1, \dots, M$  do
11:        Compute  $P_{\text{subs}}(w_j|w_i)$   $\triangleright$  Eq. 1
12:      Randomly select words to replace given WER
13:      Replace selected words with error words based on the
         probability distributions computed  $\triangleright$  Line 11

```

In Table 1, we provide several examples of the substitution errors simulated at a given WER of 30%.

3. SENTENCE EMBEDDING METHODS

The sentence embedding methods described in this section have all been shown to perform well on STS tasks [25], [26] and serve as a representative set of models to evaluate robustness to ASR errors. A brief description of each method is provided below:

Simple Unweighted Average: A common sentence embedding implementation is a computation of the arithmetic mean for all word vectors that comprise a sentence. This serves as a simple but effective baseline with pre-trained *word2vec* embeddings [2]. Additionally, averages can be computed after removing stop words which contain little semantic content (*e.g.* "is", "the", *etc.*).

Smooth Inverse Frequency (SIF): Arora *et al.* propose SIF embeddings [11], which involve two major components. First, a weighted average of the form $\frac{a}{a+p(w)}$ is computed, in which a is a scalar value (a hyperparameter, tuned to 0.001) and $p(w)$ is the probability that a word appears in a given corpus. This weighting scheme

de-emphasizes commonly used words (with high probability) and emphasizes low probability words that likely carry more semantic content. Additionally, SIF embeddings attempt to diminish the influence of semantically meaningless directions common to the whole corpus. To do so, all word vectors in a dataset are concatenated into a matrix from which the first principal component is removed from each weighted average.

Unsupervised Smooth Inverse Frequency (uSIF): Ethayarajh proposes a refinement to SIF known as uSIF, which claims improvements in many tasks (including STS) [15]. uSIF differs from SIF in that the hyperparameter a is directly computed (and not tuned), making it fully unsupervised. Additionally, the first m ($m = 5$) principal components, each weighted by the factor $\lambda_1, \dots, \lambda_m$ are subtracted for the common component removal step. Here, $\lambda_i = \frac{\sigma_i^2}{\sum_{i=1}^m \sigma_i^2}$, where σ_i is the i -th singular value of the embedding matrix.

Low-Rank Subspace: Mu *et al.* propose a unique sentence embedding in which sentences are represented by an N -dimensional subspace rather than a single vector [12]. Given word vectors of dimension d and subspace rank of N , a sentence matrix is first constructed by concatenating word vectors and has dimension $d \times N$ (we use $d = 300$ and $N = 4$). Then, principal component analysis (PCA) is performed to identify the first N principal components whose span comprise a rank- N subspace in \mathbb{R}^d . We consider this method for our simulated ASR error analysis to test whether the subspace representation is more robust to ASR errors than a vector representation.

InferSent: Conneau *et al.* developed the *InferSent* encoder that utilizes a transfer learning approach [13]. The encoder is trained with a bidirectional LSTM neural network on the Stanford Natural Language Inference (SNLI) dataset, a labeled dataset that is designed for textual entailment tasks. The embeddings learned from the NLI task are then used to perform textual similarity tasks in STS.

Computing Similarities: Sentences represented by vectors (*i.e.* averages, SIF, uSIF, *InferSent*) can be compared with *cosine similarity*, closely related to d_{ij}^S in Equation 2. Cosine similarity is given as $\text{CosSim} = 1 - d_{ij}^S = \cos \theta_{ij} = \frac{\mathbf{w}_1^T \mathbf{w}_2}{\|\mathbf{w}_1\|_2 \|\mathbf{w}_2\|_2}$. For subspace similarity, the authors in [12] suggest the analogous concept of computing the *principal angle* between the rank- N subspaces for two sentences. This can be readily obtained from the singular value decomposition. If we let the matrices $U(s_1)$ and $U(s_2)$ have columns that each contain the first N principal components for sentences s_1 and s_2 , the principal angle similarity given by:

$$\text{PrincAng}(s_1, s_2) = \sqrt{\sum_{t=1}^N \sigma_t^2} \quad (3)$$

In Equation 3, σ_t represents the t -th singular value of the product $U(s_1)^T U(s_2)$.

4. RESULTS & DISCUSSION

4.1. Robustness of Sentence Embeddings to Simulated ASR Errors

To study the effects of ASR errors on sentence embeddings, we first computed a sentence embedding for each sentence in SICK [16] and STS-benchmark [17] *dev* and *test* sets using each of the methods

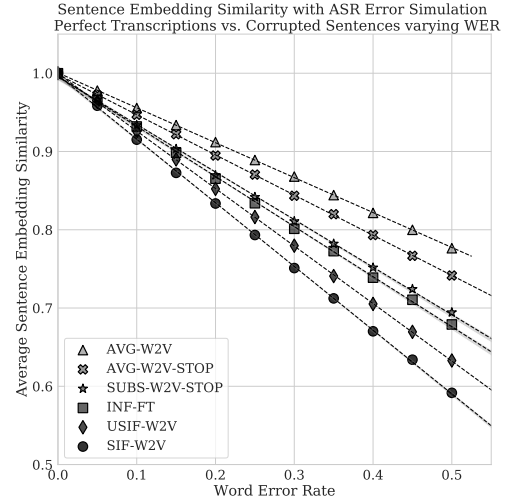


Fig. 1. Regression plots for sentence embedding methods described in Section 3 as the WER is varied from 0% to 50%. We consider averaging *word2vec* vectors (\triangle), averaging *word2vec* and removing stop words (\times), low-rank subspace representations with *word2vec* and stop-words removed (\star) [12], *InferSent* with *FastText* embeddings (\square) [13], SIF with *word2vec* [11] (\circ), and uSIF with *word2vec* (\diamond) [15].

described in Section 3. Since *GloVe* embeddings were used to generate the simulated ASR substitution errors, we used *FastText* (for *InferSent*) and *word2vec* embeddings (all other methods) to generate sentence embeddings. For each method, we corrupted the sentences in the text with a defined WER between 0% and 50% with the simulator described in Section 2. Then, each sentence in each set is compared with its corrupted counterpart using the relevant similarity metric (*i.e.* cosine or principal angle similarity).

The results are shown in Figure 1, in which all methods show a steady linear decline in average similarity between original and corrupted sentences as WER is increased. As expected, when WER is 0%, the sentence embedding similarity is equal to 1 for all methods. Simple averaging shows the least significant decline as WER is increased, *i.e.* at WER = 50% we see $\text{sim}_{\text{avg}} \approx 0.776$ for unweighted averaging and $\text{sim}_{\text{avg}} \approx 0.742$ for unweighted averaging and stop words removed. However, we see a significantly steeper decline for SIF and uSIF when WER = 50%, *i.e.* $\text{sim}_{\text{avg}} \approx 0.592$ for SIF and $\text{sim}_{\text{avg}} \approx 0.633$ for uSIF. The subspace representation and *InferSent* show a moderate decline in between these two extremes. These results are in line with our intuition, as we expect word substitution errors to have the smallest overall impact on unweighted average sentence embeddings. Also as expected, unweighted averages with stop words are more impacted by ASR errors, since stop words in the original corpus could be replaced by content words. This would lead to a greater difference between original and corrupted sentence similarity scores. SIF and uSIF are the most impacted by word substitution errors. We believe this is explained by the weighted average computation, *i.e.* if a frequent word is replaced by a less frequent word, it may have a greater impact on the overall sentence embedding. Additionally, it is likely the principal components of the embedding matrix are drastically altered by the introduced error and variance in the dataset, leading to larger differences in sen-

Sentence Embedding	STS Corpus (dev & test set)	PCC _{0%} / PCC _{10%} / PCC _{30%} ($\times 100$)	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%

Table 2. Pearson Correlation Coefficient (PCC) performance ($\times 100$) for SICK and STS-benchmark *dev* and *test* sets when WER is varied (0%, 10%, and 30%). The last column of each table shows the ratio (as a percentage) of the PCC at WER = 30% to the PCC at WER = 0% to demonstrate the robustness in STS performance of each sentence embedding to ASR errors at a high WER.

tence embedding representations after corruption and common component removal. Since the common component removal is weighted by $\lambda_i \leq 1$ for each of the i principal components in uSIF, the overall impact of the introduced variance due to ASR errors is diminished when compared to the single component removal step in SIF.

4.2. Evaluation of STS Results with Word Substitution Errors

We next compared the STS performance of the sentence embeddings on the original and corrupted corpora (with 10% and 30% WER) with the *dev* and *test* sets of SICK [16] and STS-benchmark [17]. The *Pearson Correlation Coefficient* (PCC) between the computed similarities and the annotated similarity scores in the corpora is the standard metric by which we evaluate STS performance of a given method. The results are seen in Table 2 and Figure 2.

On the original sentences, simple unweighted averaging provides a strong benchmark for STS tasks on both corpora, with nearly equivalent results when stop words are removed. In most cases, the weighted average and de-noising provided by SIF and uSIF improve upon the results of unweighted averages, with both methods displaying near-identical performance. The subspace results are somewhat inconclusive, as they show a slight improvement over averages, SIF, and uSIF on STS-benchmark but a decrease in performance on SICK. The authors in [12] chose $N = 4$ empirically as the subspace rank, based on a variety of corpora which comprise the STS-benchmark set. It is possible that the absolute performance of the subspace sentence embedding can be improved by tuning the fixed subspace rank for SICK as well. Unsurprisingly, *InferSent* is consistently the strongest performer, likely due to its supervised training on the SNLI corpus.

When, ASR errors are introduced, the STS performance for each method changes significantly, as evidenced by the results in Table 2. Though the simple averages were least impacted with the introduction of ASR errors (Section 4.1), they perform worst among the methods tested on STS tasks with a high WER. On the other hand, SIF and uSIF embeddings were most impacted by ASR errors but perform among the best in STS when the WER is high. Again, we suspect this is due to the common component removal steps in SIF and uSIF, which effectively act as de-noising steps removing some of the additional variance in the embedding matrix due to substitution errors. Since SIF and uSIF display near-identical STS performance

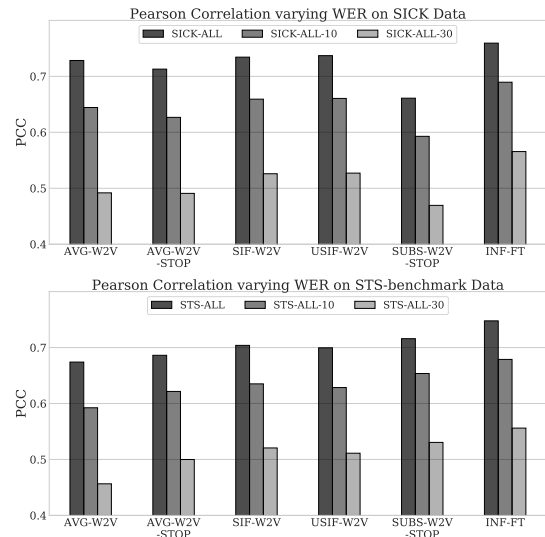


Fig. 2. Graphical depiction of the STS performance of various sentence embeddings with simulated word substitution error, see Table 2

across both corpora, we think uSIF may be a slightly better choice due to its increased robustness to ASR errors. Also, as suspected, we see that the subspace embeddings show increased STS performance robustness to word substitution errors when compared to averages if we consider the PCC ratio between high WER (30%) and original sentences. Subspace embeddings slightly outperform SIF and uSIF on STS-benchmark and slightly under-perform SIF and uSIF on SICK by the same metric. Again, *InferSent* not only shows the best absolute performance on the original sentences, but shows the best performance with a high WER rate as well.

5. CONCLUSION

In this paper, we introduced a simulator that automates word substitution errors (given a WER) on perfectly transcribed corpora to simulate ASR-plausible errors, considering both phonemic and semantic similarities between words. We then used the simulator to intentionally corrupt standard corpora used for textual similarity tasks (SICK [18] and STS-benchmark [17]). From this, we were able to evaluate the impact that word substitution errors may have on some of the most recently developed techniques for sentence embeddings. We also evaluated the STS performance of each of these sentence embedding methods after introducing substitution errors with our simulator. We found several interesting results. For example, average sentence embeddings perform well for perfectly transcribed text, but show poorer STS performance when errors are introduced if compared to more advanced methods. On the other hand, pre-trained encoders, such as *InferSent* not only show state-of-the-art performance on STS tasks with perfectly transcribed text, but also seem to show increased robustness to error for STS performance. If it is not possible to use an encoder like *InferSent*, the weighted average and smoothing provided by SIF/uSIF or the low-rank subspace representation by Mu *et al.* [12] seem to be reasonable improvements over simple averages when it comes to STS performance for high-WER transcriptions.

6. REFERENCES

- [1] Thomas K. Landauer and Susan T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.
- [2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [3] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global Vectors for Word Representation," 2014, pp. 1532–1543, Association for Computational Linguistics.
- [4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," *arXiv:1607.04606 [cs]*, July 2016.
- [5] Gamal Bohouta and Veton Këpuska, "Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx)," *Int. Journal of Engineering Research and Application*, vol. 2248-9622, pp. 20–24, Mar. 2017.
- [6] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee, "Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension," in *Inter-speech 2018*. Sept. 2018, pp. 3459–3463, ISCA.
- [7] Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, and Yannick Estève, "Simulating ASR errors for training SLU systems," in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018, p. 7, European Language Resources Association.
- [8] Matthew N. Stuttle, Jason D. Williams, and Steve Young, "A framework for dialogue data collection with a simulated ASR channel," in *Eighth International Conference on Spoken Language Processing*, 2004.
- [9] Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, and Gary Geunbae Lee, "An integrated dialog simulation technique for evaluating spoken dialog systems," in *Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, 2008, pp. 9–16.
- [10] Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger, "From Word Embeddings To Document Distances," p. 10.
- [11] 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.
- [12] Jiaqi Mu, Suma Bhat, and Pramod Viswanath, "Representing Sentences as Low-Rank Subspaces," *arXiv:1704.05358 [cs]*, Apr. 2017.
- [13] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *arXiv:1705.02364 [cs]*, May 2017.
- [14] Matteo Pagliardini, Prakhhar Gupta, and Martin Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," *arXiv:1703.02507 [cs]*, Mar. 2017.
- [15] 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.
- [16] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli, "SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 1–8, Association for Computational Linguistics.
- [17] 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.
- [18] 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.
- [19] Nathan C. Sanders and Steven B. Chin, "Phonological Distance Measures*," *Journal of Quantitative Linguistics*, vol. 16, no. 1, pp. 96–114, Feb. 2009.
- [20] 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.
- [21] Kathleen Currie Hall, Blake Allen, Michael Fry, Scott Mackie, and Michael McAuliffe, "Phonological CorpusTools," in *14th Conference for Laboratory Phonology*, Tokyo, Japan, 2015.
- [22] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, pp. 707, Feb. 1966.
- [23] Robert L. Weide, "The CMU pronouncing dictionary," URL: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>, 1998.
- [24] Bruce Hayes, "Introductory Phonology," *Blackwell Textbooks in Linguistics*, 2009.
- [25] Yves Piersmen, "Comparing Sentence Similarity Methods," Feb. 2018.
- [26] Christian S. Perone, Roberto Silveira, and Thomas S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *arXiv:1806.06259 [cs]*, June 2018.