Empirical Study of Sentence Embeddings for English Sentences Quality Assessment

Pablo Rivas and Marcus Zimmermann
Department of Computer Science, School of Computer Science and Mathematics
Marist College, 3399 North Road Poughkeepsie, New York 12601, United States

Abstract—*Novel deep learning and machine translation techniques have greatly advanced the field of computational linguistics, enabling us to find meaningful latent spaces for text analysis. While several embedding techniques exist for words, sentences, and entire documents, the potential applications are still being explored. In this paper we present the impact of top-performing sentence embedding methodologies on the accuracy of a neural model trained to assess the quality of English sentences. We focus our efforts in the methodologies called Language Agnostic SEntence Representation (LASER), Sentence to Vector (S2V), and Universal Sentence Encoder (USE) to observe their ability to capture information related to sentence quality. Our study suggests that these state-of-the-art sentence embeddings are unable to capture sufficient information regarding sentence correctness and quality in the English language.*

Keywords: sentence embeddings; sentence quality; neural networks; sentence encoding

1. Introduction

The last few years have seen remarkable improvement in sentence embedding algorithms intended to capture linguistic properties, with embeddings successfully applied to document analysis and categorization, as well as document sentiment and tone analysis. Some of the most notable efforts include Google's Universal Sentence Encoder (USE) for English [1], which uses a deep averaging neural network to encode sentences, and transfer learning at the word and sentence level to achieve solid performance on various NLP tasks. Another successful approach is Sent2Vec [2], which executes both supervised and unsupervised learning over compositional n-gram features to learn sentence embeddings. Facebook also produced an embedding methodology, Language-Agnostic SEntence Representations (LASER) [3]. LASER uses encoder-decoder architectures inspired by neural machine translation models, producing sentence embeddings that are language agnostic. Recent studies [4] have shown that these three embeddings outperform others in assessments concerning the preservation of linguistic properties in sentence representations. These research developments were considered in our group's most recent study, which is concerned with assessing the quality of English sentences according to five specific rules [5]. It is worth noting that these rules, such as the proper application of subjects and verbs in a sentence, or the stylistic preference for brevity, can be highly subjective. However, we believe these rules capture fundamental elements of style and grammar, serving as a strong indicator for proper writing and technique.

Our study evaluates the ability of these top-performing sentence embeddings to capture sentence correctness and quality. Using a dataset of English sentences that contains mixed levels of writing proficiency, we trained a neural network to classify properly- and poorly-constructed sentences, documenting the impact of each embedding methodology on the overall performance of the network. Our study indicates that these state-of-the-art methodologies, USE, Sent2Vec, and LASER, are unable to embed enough information regarding sentence correctness and quality.

The remainder of this paper is organized as follows: Section 2 introduces the current state of sentence embedding in the research community, as well as the distinct methodologies of USE, Sent2Vec, and LASER. Section 3 explains the implementation of these sentence embeddings in the context of our neural model. Section 4 explains the impact of using these sentence embeddings on the accuracy of our neural model. Finally, Section 5 concludes our paper with a discussion of findings and plans for future work.

2. Embedding Methodologies

Sentence embedding is critical to the success of NLP applications, but few individuals understand exactly how these embeddings, and the linguistic properties they capture, actually impact downstream tasks. Furthermore, different embedding schemes can produce completely different representations. These representations may be similar insofar as having the ability to conduct limited analysis or categorization of sentences. But the actual linguistic properties captured in these vectorized representations may be quite different, making a lot of experimentation with embeddings blind trial and error. Now that strong performance baselines have been established for many NLP applications [6], researchers and developers have increasingly turned their attention to probing tasks that assess linguistic properties captured in different sentence embeddings. Analyzing tense, clause dependency, and other linguistic properties is a great way to determine the strength of a representation, and it gives us a better understanding of how exactly these top-performing embedding schemes work [7]. In the following paragraphs, we will examine the methodologies of USE, Sent2Vec, and LASER, all of which are considered to be state of the art [4].

Google's Universal Sentence Encoder (USE) has two implementations, each modeled to achieve a different design goal. One makes use of a transformer architecture, targeting high accuracy at the cost of additional complexity and resource consumption, and the other is formulated as a deep averaging network (DAN), targeting speed and quick inference at the cost of accuracy [1]. For our research, we used the DAN encoder, which takes a tokenized string as input and produces a 512-dimensional sentence embedding. This output embedding is then averaged with other embeddings produced by the encoder and passed into a feedforward deep neural network. Averaging may sacrifice accuracy, but the DAN encoder still manages to achieve strong performance baselines on a variety of NLP tasks, and even matches or outperforms the transformer encoder in some unique cases. Furthermore, the DAN encoder operates efficiently, as expected. It's compute and memory usage is $\mathcal{O}(n)$, whereas the transformer encoder is $\mathcal{O}(n^2)$, making it a strong choice for an application's encoder.

Sent2Vec can be thought of as a natural extension of tools like FastText and Word2Vec; however, the goal is to vectorize word sequences, not just words. With Sent2Vec, sentence embeddings are produced by averaging the source word embeddings of the sentence's constituent words, with source embeddings including not only unigrams but also n-grams present in the sentence [2]. Depending on the implementation, this results in either a 600- or 700-dimensional embedding. This approach was inspired by simpler models like matrix factorization in an effort to exploit the fact that they are computationally inexpensive, allowing them to efficiently tackle larger sets of data. With this design goal in mind, Sent2Vec successfully achieves $\mathcal{O}(1)$ vector operations per word processed, making it scalable, efficient, and therefore another strong, practical encoder.

Lastly, Facebook's Language Agnostic SEntence Representation (LASER) is inspired by the encoder-decoder architectures of neural machine translation models. LASER uses a single, shared encoder that accepts sentences in any of 93 languages as input and maps them to a point in high-dimensional space. The final representation is a 1,024-dimensional vector, with the goal being a universal language capable of representing sentences from different languages, though identical in meaning, in a vectorized representation that is virtually the same. This vectorized representation is then used as input for the decoder, along with a specification regarding the desired output language. LASER was recently shown to embed more linguistic information than its competitors [4], outperforming them on 17 out of 22 SentEval tasks. With the highest dimensional representation, these embeddings naturally take a little longer to process. However, LASER is still capable of processing roughly 2000 sentences per second on a GPU, making it another excellent embedding tool.

3. Models Implementation and Experiment Design

The sentence embedding methodologies discussed in the previous section are applied in a dataset that aims to assess the quality of a sentence using non-trivial rules in the English language.

3.1 Rules

There are well-known rules that make a sentence *a good sentence*. In our research [5], we focused on the following five rules:

1) **Subjects**. The subject must be the main character, not actions expressed as abstract nouns.
2) **Verbs**. The important actions in the sentence should be verbs, not abstract nouns.
3) **Introductory Phrases**. Introductory phrases in a sentence (if any) should follow rules 1 and 2, and should not be too long; around five words is acceptable.
4) **Nouns**. Strings of consecutive nouns (three or more) should be avoided to preserve sentence clarity.
5) **Conciseness**. Words that mean little or nothing, words that repeat the meaning of other words, words implied by other words, should all be avoided.

These rules have been used by our experts to evaluate sentences and create a dataset of sentences with these quality markers.

3.2 Dataset

Due to the natural complexities of the English language, the rules previously mentioned are nearly impossible to define in a precise manner without falling into endless exceptions; for this reason we created a dataset to train machine learning algorithms to model such non-trivial rules. The data comes from college students writing after the removal of all personal identification data. Our experts observed a sentence and evaluated it on the five rules. The average length of the sentences in the dataset is 110 words, and the vocabulary size is 73,140. The sentences can be visualized using tools

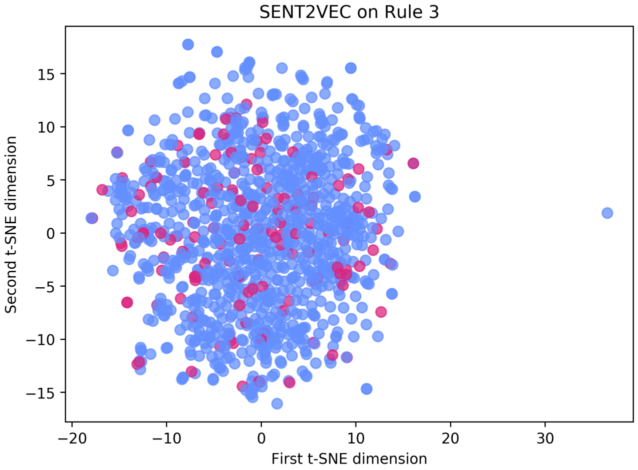

Fig. 1: Two-dimensional t-SNE visualization of Sent2Vec embeddings. Clear dots are correct sentences and dark dots are sentences that violate rule three.

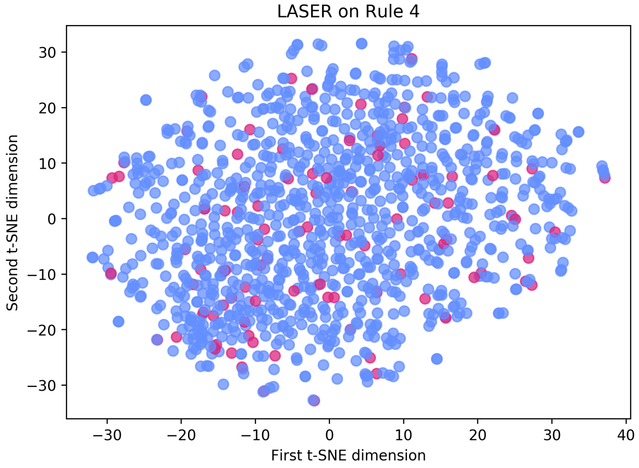

Fig. 2: Two-dimensional t-SNE visualization of LASER sentence embeddings. Clear dots are correct sentences and dark dots are sentences that violate rule four.

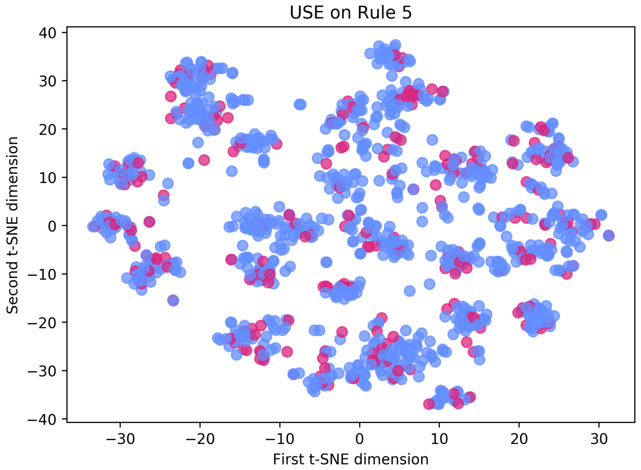

Fig. 3: Two-dimensional t-SNE visualization of USE sentence embeddings. Clear dots are correct sentences and dark dots are sentences that violate rule five.

such as t-SNE over the embeddings methodologies discussed before.

3.3 Sentence Embeddings and Visualization

A popular visualization tool for high-dimensional data is t-SNE [8]. This algorithm can find optimal ways to display data in low-dimensional spaces. It is typically used to display data in two dimensions to see if there are clear clusters of data or separability of labeled groups.

Here we used t-SNE to display sentences. First, Fig. 1 shows the sentence embeddings produced using Sent2Vec projected down to two dimensions. The figure shows no discernible separation between sentences that do or do not violate the third rule on *Introductory Phrases*.

Similarly, Fig. 2 depicts the embedding space produced by the LASER methodology encoded in two dimensions. As can be seen, there are no obvious groupings of correct or incorrect sentences on rule number four about *Nouns*.

However, Fig. 3 shows distinct data groups when using the USE algorithm. This two-dimensional representation displays several groups. Unfortunately, these groups are not directly related to sentence quality. It is very likely that these groups are related to topical sentence information; but this remains a conjecture at this point, and will require further topical classification of our sentence dataset, which goes beyond the scope of this research. The sentences in Fig. 3 correspond to the assessment of the fifth rule about *Conciseness*.

Note that although sentences can be visualized in two dimensions using t-SNE, the full high-dimensional latent space is the object of our study. We discuss this next in the context of dense neural networks.

3.4 Dense Neural Networks for Quality Assessment

The sentence embeddings vary in methodology and in the dimensions into which they encode. USE encodes into 512 dimensions; LASER encodes into 1,024 dimensions; and Sent2Vec can encode into 600 or 700 dimensions depending on which model is used. Fig. 4 on the left depicts a variable latent space corresponding to any of the embedding methodologies.

The right hand side of Fig. 4 depicts three dense layers

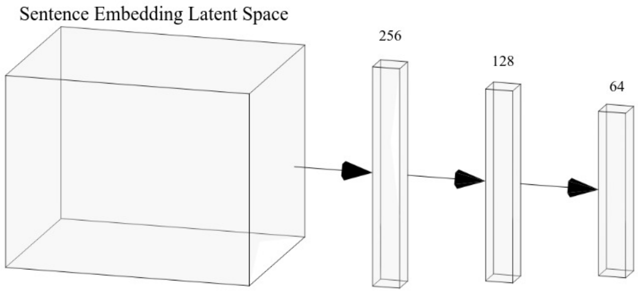

Fig. 4: Sentence embeddings are connected to dense layers. The number of layers is treated as a hyper-parameter and is chosen through a cross-validated grid search.

of fully connected neural networks, each with 256, 128, and 64 neurons, respectively. These three simple layers are treated as a hyper-parameter that needs to be optimized, meaning that the best number of layers is determined for each methodology. As will be seen in the next section, in the simplest case the dense network will have a single layer of 64 neurons, or in the more complex case it will have the complete three layers as depicted in the figure.

The experiments consist of embedding our sentence dataset using each different methodology, *i.e.*, LASER, USE, and Sent2Vec. Then, we find the dense architecture that will produce the best results using 6-fold cross validation. Once the best architecture is determined, the balanced accuracy is estimated using 10-fold cross validation.

4. Evaluation Results and Discussion

Table 1 shows the results of applying 10-fold cross validation to the problem of evaluating the balanced accuracy rate for each of the methodologies that are the subjects of this study.

In all of our experiments, the accuracy rate reported corresponds to the *Balanced Accuracy* defined as follows:

$$\text{Acc} = \left(\frac{TP}{P} + \frac{TN}{N} \right) /2$$

where TP is the count of predicted true positives, P is the count of positives, TN is the count of predicted true negatives, and N is the count of negatives.

From Table 1 we can observe all the variations of Sent2Vec depending on the dataset on which the embedding model was trained. The best scores are shown in bold font. For rules one and four, the best embedding model was Facebook's LASER embedding method. For rules two and three, Sent2Vec is the best methodology using Toronto Books Unigrams and Twitter Bigrams data, respectively. For rule five, Google's USE reported the highest performance.

Looking at the aggregated statistics across rules and methodologies in Table 1, we can observe that the rule that produces the highest accuracy is the fifth rule on *Conciseness*. On the other hand, the top three methodologies for embedding across all rules are, Sent2Vec (trained on Twitter Bigrams), USE, and LASER.

The next logical step in the experimentation process was to test combinations of the top methodologies in groups of two. The combination is achieved using a simple concatenation of embeddings. For example, if the embedding produced by USE and LASER for the i-th sentence is defined as $\mathbf{x}_i^{\text{USE}} \in \mathbb{R}^{512}$ and $\mathbf{x}_i^{\text{LASER}} \in \mathbb{R}^{1024}$, respectively, then we can define the concatenation as follows:

$$g_i^{(3)} = [\mathbf{x}_i^{\text{USE}} \quad \mathbf{x}_i^{\text{LASER}}]^T$$

where g_3 is simply a reference to the third group formed. All the groups formed are shown in Table 2.

From Table 2, we can see that the best pair of methodologies combined is USE and LASER when considering the aggregated statistic. With respect to the rule that yields the best performance, it can be seen that the fifth rule on *Conciseness* is the one that is classified better. Notice that the same procedure of selection of the best dense architecture is followed, and the results reported are cross-validated.

Table 3 presents the results of the final experiment in which we combine the top three methods into a single latent vector. The purpose is to examine the capabilities of this combination to perform a quality assessment. From the table we can see that the balanced accuracy is the highest for rule five. It can be easily seen that all the results are very close to random change predictions, that is, balance accuracy rates close to 0.5. The only rule that is modeled slightly better than random chance is the rule of *Conciseness*. It can be argued that longer sentences, which may have greater probabilities of being labeled as violating the *Conciseness* rule, are more easily detected than the other quality rules.

While these embedding methodologies have been proven to preserve lexical information [4] based on the empirical evidence shown here, we can make the claim that these state-of-the-art methodologies are unable to capture and encode enough information related to the quality of a given sentence.

Furthermore, simple combinations of the top performers are still not able to perform significantly better than random chance, which is problematic. This suggests that the sentence quality assessment of non-trivial and subjective rules on the English language is a problem that has not been solved. More research is needed in the embedding of sentences for purposes of quality assessment [5].

We want to acknowledge that the top performers we studied here, USE [1], LASER [3], and Sent2Vec [2], have excellent performance in many other language-related

Table 1: Cross-validated accuracy of individual methodologies on the problem of sentence quality assessment.

Methodology		Rule 1		Rule 2		Rule 3		Rule 4		Rule 5		**Average**	
Algo	Dataset	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ
S2Vec	Wiki Unigrams	0.484	0.03	0.499	0.04	0.487	0.03	0.489	0.02	0.533	0.05	0.498	0.03
	Wiki Bigrams	0.509	0.05	0.497	0.02	0.503	0.04	0.499	0.04	0.515	0.05	0.505	0.04
	Toronto Big.	0.502	0.02	0.489	0.03	0.504	0.03	0.486	0.01	0.522	0.05	0.501	0.03
	Toronto Unig.	0.498	0.04	**0.516**	0.05	0.497	0.04	0.499	0.02	0.519	0.04	0.506	0.04
	Twitter Unig.	0.490	0.02	0.504	0.04	0.502	0.04	0.502	0.02	0.531	0.04	0.506	0.03
	Twitter Big.	0.498	0.01	0.505	0.03	**0.509**	0.04	0.483	0.03	0.537	0.04	**0.506**	0.03
USE	Web-Sources	0.508	0.03	0.488	0.03	0.501	0.02	0.499	0.04	**0.546**	0.08	**0.509**	0.04
Laser	UN Corpus	**0.528**	0.05	0.510	0.05	0.487	0.02	**0.534**	0.05	0.538	0.04	**0.520**	0.04
	Rule Average	0.502	0.03	0.501	0.03	0.499	0.03	0.499	0.03	**0.530**	0.05		

Table 2: Cross-validated accuracy of grouped methodologies by pairs for the problem of sentence quality assessment.

Methodology		Rule 1		Rule 2		Rule 3		Rule 4		Rule 5		**Average**	
Group	Algorithms	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ
$g^{(1)}$	S2Vec + USE	0.503	0.02	0.498	0.03	0.482	0.03	0.509	0.02	0.529	0.03	0.504	0.02
$g^{(2)}$	S2Vec + Laser	0.499	0.01	**0.517**	0.04	0.498	0.05	0.484	0.03	**0.545**	0.04	0.509	0.02
$g^{(3)}$	USE + Laser	**0.528**	0.04	0.484	0.04	**0.498**	0.01	**0.514**	0.03	0.532	0.06	**0.511**	0.02
	Rule Average	0.510	0.03	0.500	0.04	0.493	0.03	0.502	0.03	**0.535**	0.05		

Table 3: Cross-validated accuracy of the top three methodologies combined for sentence quality assessment.

Methodology	Rule 1		Rule 2		Rule 3		Rule 4		Rule 5		**Average**	
Algorithms	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ	Acc	σ
S2Vec + USE + Laser	0.514	0.04	0.514	0.03	0.499	0.05	0.503	0.03	**0.525**	0.05	**0.511**	0.01

applications. This is mainly because they have been trained using traditional and novel machine translation techniques. Machine translation has paved the way to solve other exciting problems in sentence analysis [9], [3], [10], [11], [12], [13]. However, a recent study [14] has shown that you can do many things when embedding a sentence into a vector, but many of these are not related to its quality.

In machine translation tasks, one usually tries to convey meaning and important aspects of the sentence; however, sentence quality is language-dependent and quality indicators, such as the ones we studied here, tend to be discarded. Nonetheless, according to Conneau *et.al.* [14], even machine translation-inspired methods are able to preserve sentence length information based on the number of words, which is congruent with our findings in Tables 1, 2, and 3, where the Rule 5 yields the best average accuracy. Since Rule 5 is related to conciseness, one can make the case that conciseness is directly related to sentence length; this would explain such results. For completeness, Fig. 5 shows the t-SNE-induced two-dimensional representation of the combined latent spaces for Rule 5, and the corresponding cross-validated receiver operating characteristic (ROC) curve analysis in Fig. 6. The corresponding cross-validated area under the curve (AUC) is of 0.58, which is slightly above random chance.

5. Conclusions

We have examined the top sentence embedding methodologies at the time of writing this paper with the purpose of assessing the quality of English sentences. Quality is measured according to five non-trivial and rather subjective rules modeled using machine learning. The top methodologies, Sent2Vec, USE, and LASER, are used to produce sentence embeddings and then classify sentences that have been tagged by experts that formed a dataset. The classification strategy uses the embeddings to train dense neural networks of different sizes to establish the baseline accuracy of each embedding strategy. Then groups of the top performing embeddings are combined to analyze the combined

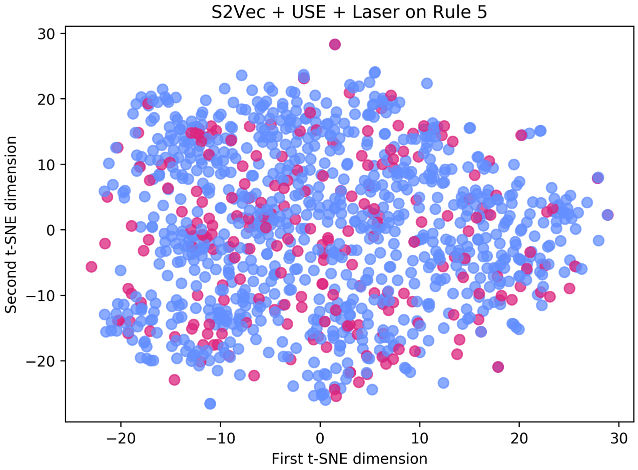

Fig. 5: Two-dimensional t-SNE representation on rule five.

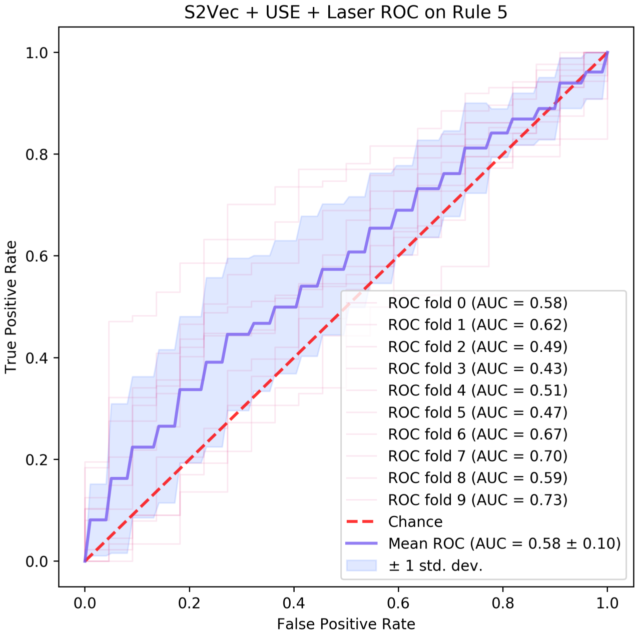

Fig. 6: Cross-validated ROC analysis on rule five using the three top sentence embedding methods.

performance. Experimental results suggest that these top performers are unable to embed enough information related to sentence quality, showing performances not significantly higher than random chance.

Further work will include non-linear combinations of these embeddings and the reproduction of the original embeddings, but trained specifically for the purpose of quality assessment rather than machine-translation purposes. Similarly, we will reproduce the models presented in the top performers and implement them for the sentence quality assessment rather than machine translation.

Code to reproduce our experiments can be found in this Google Colaboratory: `http://marist.ai/sent-emb`

Acknowledgements

We want to thank the subject matter experts at Marist College (MC), including professors and students who tagged sentences forming our initial dataset. This work was supported in part by the New York State Cloud Computing and Analytics Center, and by the VPAA's office at MC.

References

[1] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal sentence encoder for english," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 169–174.

[2] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 528–540.

[3] H. Schwenk and M. Douze, "Learning joint multilingual sentence representations with neural machine translation," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017, pp. 157–167.

[4] K. Krasnowska-Kieraś and A. Wróblewska, "Empirical linguistic study of sentence embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5729–5739.

[5] P. Rivas, "Modeling five sentence quality representations by finding latent spaces produced with deep long short-memory models," in *Proceedings of the 2019 ACL Workshop on Widening NLP*, 2019, pp. 24–26.

[6] C. S. Perone, R. Silveira, and T. S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *arXiv preprint arXiv:1806.06259*, 2018.

[7] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[8] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[9] J. Wieting and K. Gimpel, "Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, 2018, pp. 451–462.

[10] Q. Cao and D. Xiong, "Encoding gated translation memory into neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3042–3047.

[11] M. Guo, Q. Shen, Y. Yang, H. Ge, D. Cer, G. H. Abrego, K. Stevens, N. Constant, Y.-H. Sung, B. Strope, *et al.*, "Effective parallel corpus mining using bilingual sentence embeddings," *arXiv preprint arXiv:1807.11906*, 2018.

[12] H. Aldarmaki and M. Diab, "Scalable cross-lingual transfer of neural sentence embeddings," in *Proc. of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*, 2019, pp. 51–60.

[13] V. Chaudhary, Y. Tang, F. Guzmán, H. Schwenk, and P. Koehn, "Low-resource corpus filtering using multilingual sentence embeddings," *arXiv preprint arXiv:1906.08885*, 2019.

[14] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," in *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, vol. 1. Association for Computational Linguistics, 2018, pp. 2126–2136.