YNU-HPCC at SemEval-2022 Task 2: Representing Multilingual Idiomaticity based on Contrastive Learning

Kuanghong Liu, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
liukh@mail.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper will present the methods[1] we use as the YNU-HPCC team in the SemEval-2022 Task 2, Multilingual Idiomaticity Detection and Sentence Embedding. We are involved in two subtasks, including four settings. In subtask B of sentence representation, we used novel approaches with ideas of contrastive learning to optimize model, where method of CoSENT was used in the pre-train setting, and triplet loss and multiple negatives ranking loss functions in fine-tune setting. We had achieved very competitive results on the final released test datasets. However, for subtask A of idiomaticity detection, we simply did a few explorations and experiments based on the xlm-RoBERTa model. Sentence concatenated with additional MWE as inputs did well in a one-shot setting. Sentences containing context had a poor performance on final released test data in zero-shot setting even if we attempted to extract effective information from CLS tokens of hidden layers.

1 Introduction

Meaning of sentence could be captured by compositionality of word representations. However, there widely exists potentially idiomatic phrases in different languages, which are multiword expressions (MWEs) with idiomatic and literal meanings. Therefore, representation of idiomatic phrases is not directly compositional. A previous study has shown that the representation of idiomatic phrases by contextual models, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and some of its variants, are not accurate(Garcia et al., 2021). It will be challenging to represent the MWEs correctly in the downstream tasks (Tayyar Madabushi et al., 2021). SemEval-2022 Task 2, Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022), involves English (EN), Portuguese (PT), and Galician (GL). This task includes two subtasks and each subtask contains two settings:

- Subtask A: Determining whether a sentence contains an idiomatic expression. This is a binary classification task with two settings of zero-shot and one-shot. Zero-shot setting means that idiomatic phrases (MWEs) in training examples are completely disjoint to those in development, evaluation, and test sets. In a one-shot setting, MWEs appearing in development and test sets include in the training sentences.

- Subtask B: Outputing the correct Semantic Text Similarity (STS) scores between sentence pairs whether or not either sentence contains an idiomatic expression. The STS scores represent semantical similarity between two sentences ranging from 0 (least similar) to 1 (most similar). This is a regressive task with two setting of pre-train and fine-tune. In the pre-train setting, models require to be trained on any semantic text similarity dataset without idiom. The fine-tune setting should use provided training sets included MWEs.

The remainder of this paper is organized as follows. In Section 2, we describe the structures of model and system. The details about data and implementation and comparative results are presented in Section 3. Finally, a conclusion is drawn in Section 4.

2 System Overview

2.1 Subtask A: Idiomaticity Detection

This is a binary classification task that requires classifying sentences into either *Idiomatic* or *Non-idiomatic*. mBERT (BERT multilingual base model) was used as a pre-trained model in the baseline method. It is a masked language models pre-

[1] The code of this paper is available at: https://github.com/lkh-meredith/SemEval2022_Task2_YNU-HPCC

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 211 - 216
July 14-15, 2022 ©2022 Association for Computational Linguistics

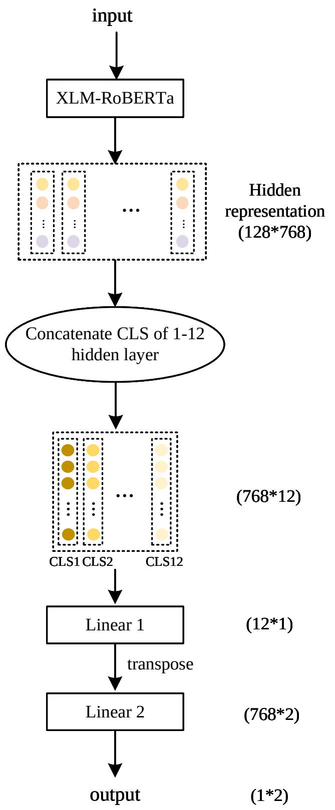

Figure 1: Model architecture of zero-shot setting

trained on the top 104 languages with Wikipedia, which is able to map multilingual representation to the same semantic space, but unable to master cross-lingual information. XLM-RoBERTa (Conneau et al., 2020) was a cross-lingual language model used for our experiment in this subtask, which combined XLM (CONNEAU and Lample, 2019) and RoBERTa (Liu et al., 2019) pre-trained on the 2.5TB of filtered CommonCrawl data containing 100 languages. It is good for the scarce language corpus that could use information learned from larger corpus of other languages. Figure 1 is process of zero-shot setting. Linear 1 layer extracted effective information from concatenated CLS tokens from 1-12 hidden layer of model's output. In the one-shot setting, we simply extracted CLS from the last hidden layer of model's output to classify.

2.2 Subtask B: Sentence Representation

In the pre-train setting , the methodology of baseline was that a sentence transformer model was created by mBERT model adding MWE tokens and training it. As shown in Figure 2 (a), it based SBERT (Reimers and Gurevych, 2019) architecture with the regression objective function, which is a good way of breaking compositionality of idiomatic phrase (Tayyar Madabushi et al., 2021). Figure 2 (b) illustrated one of the methods that we took, which was a siamese network structure of SBERT with classification objective function. Vectors $u, v, \parallel u - v \parallel$ was concatenated as a feature, and $\parallel u - v \parallel$ could play a crucial role in determining if two sentences were similar. Figure 2 (c) illustrates a new method of optimizing cosine similarity, CoSENT (Cosine Sentence), which was proposed by Jianlin Su in his blog post[2]. In the method (c), sentences in a batch are composed of sentences pairs, where two sentences that belong to one sentence pair are adjacent, so they can not be shuffled. The most important part of it was that a loss function based on contrastive learning was designed to maintain training and prediction consistency. CoSENT loss function defined as follows,

$$\log\left[1+\sum_{\mathrm{sim(i,j)>sim(m,n)}} e^{\lambda(\cos(\mathrm{u_m},\mathrm{u_n})-\cos(\mathrm{u_i},\mathrm{u_j}))}\right] \qquad (1)$$

where (i, j) and (m, n) are sentence pairs, u_i, u_j, u_m, u_n are sentence embeddings. λ is a hyper-parameter of 20 in our experiment. $e^{\lambda(\cos(u_m,u_n)-\cos(u_i,u_j))}$ is added when label of (i, j) is greater than (m, n), so $\cos(u_i, u_j) > \cos(u_m, u_n)$ is expected in the loss function.

The methodology of baseline used for the fine-tune setting is similar to the pre-train setting: create a sentence transformer model with mBERT adding MWE tokens. This sentence transformer firstly output scores for some of fine-tune data that had no scores (details about data in section 3.1) and trained on them. The questions about the method is that: 1) It is not good to calculate similarity directly between sentence vectors generated by pre-trained model without fine-tuing, which generate static labels and may not accurate. 2) mBERT are not trained on the parallel data, so their vector space across languages are not aligned, which may result in poor results on other languanges. We chose a sentence transformer, distiluse-base-mutilingual-cased (Reimers and Gurevych, 2019), provided in Hugging Face models hub[3], which had been demonstrated to generate good sentence embeddings in

[2] https://kexue.fm/archives/8847

[3] distiluse-base-mutilingual-cased-v1: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1
distiluse-base-mutilingual-cased-v2: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

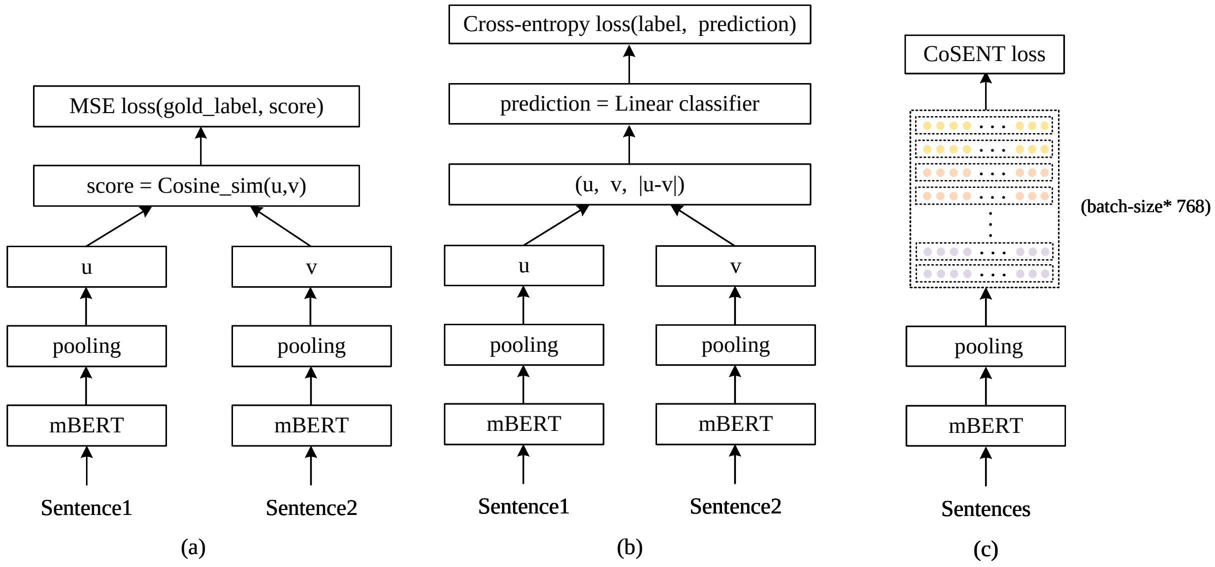

Figure 2: The methods of subtask B pre-train setting

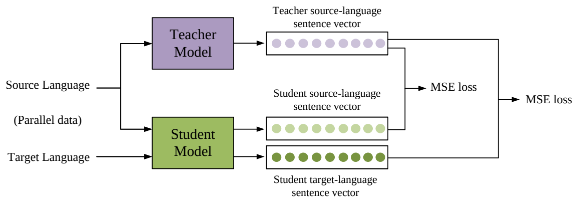
Teacher Model
Teacher source-language sentence vector
Source Language
(Parallel data)
Target Language
Student Model
Student source-language sentence vector
MSE loss
MSE loss
Student target-language sentence vector

Figure 3: The method of making monolingual sentence embeddings multilingual using knowledge distillation

multilingual language and been evaluated in task of multilingual sematic textual similarity. The the difference of its two version, v1 and v2, is that mUSE sentence encoder supports 15 languages and v2 supports 50 including GL respectively. The training approach was achieved by using knowledge distillation. It is able to extend sentence embedding from source language to target ones by sentence pairs of translation (Reimers and Gurevych, 2020). The training procedure of this pre-trained model is shown in Figure 3 using mUSE (Yang et al., 2020) as teacher model and distilmBERT (a distilled version of the mBERT) as student model. Mean pooling of outputs are as sentence vector and minimize the mean-squared loss. Therefore, this model applies to semantic similarity task, and it will be removed last dense layer was as extractor of feature in our experiment.

Triplet loss and multiple negatives ranking loss functions were used to fine-tune. Triplet loss function was used to fine-tune model so that the distance of correct sentence pairs should be closer than incorrect sentence pairs. It computed as follows:

$$\max(\| \, anchor - pos \, \| - \| \, anchor - neg \, \| \\ + margin, 0) \tag{2}$$

The multiple negatives ranking loss (Henderson et al., 2017) was used to implement and optimize, as shown in Figure 4. $(S_1, S'_1) \ldots (S_n, S'_n)$ were positive pairs. The matrix X represented cosine similarity between sentence pairs in a batch. For a batch of size n, there would be n targets ($y = (0, 1, \ldots, n-1)$) treated as labels to represent position of positive pairs in the matrix. With increasing batch sizes, the performance usually is better. The approximated mean negative log probability of data was realized to calculated by cross entropy loss in Pytorch:

$$\begin{aligned} & -\frac{1}{n} \sum_{i=0}^{n-1} \log \frac{e^{X_{i,y_i}}}{\sum\limits_{j=0}^{n-1} e^{X_{i,y_j}}} \\ & = -\frac{1}{n} \sum_{i=0}^{n-1} \left[X_{i,y_i} - \log \left(\sum_{j=0}^{n-1} e^{X_{i,y_j}} \right) \right] \end{aligned} \tag{3}$$

3 Experiments

3.1 Dataset

There includes EN, PT, and GL in datasets. GL is only provided in test dataset and one-shot training data, aiming to test models' ability to transfer learning across languages. Besides, Modern Galician (GL) is part of the West Iberian languages group, which belongs to a family of Romance languages including the Portuguese. A brief intro-

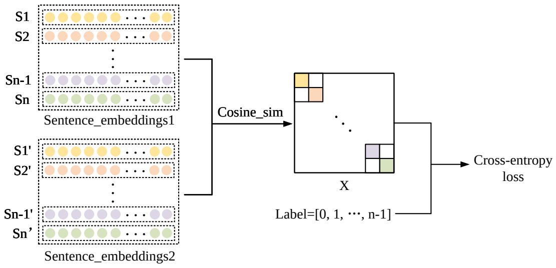

Figure 4: Multiple negatives ranking loss used for positive pairs

duction about datasets describes as follows, and more details are in the task description paper (Tayyar Madabushi et al., 2022).

In the subtask A, training data provided includes zero-shot and one-shot data. A label of 0 indicates *Idiomatic* and a label of 1 indicates *Non-idiomatic*. Zero-shot training data combined with the context, sentences preceding and succeeding the one containing the idioms, are used as training sentences in the zero-shot setting. In the one-shot setting, both the zero-shot and one-shot data are used to train, which exclude the context but add the MWEs as an additional feature. Sentence was concatenated to MWE and separated them by SEP.

Train split of STSBenchmark datasets in English and ASSIN2 datasets in Portuguese are as training dataset in the subtask B pre-train setting. Development and test datasets consist of sentence pairs, some of whom contains idiomaticity and other has no idiomaticity, which replaced by non-idiomatic paraphrases. They will be computed a score (cosine similarity) after model's outputing.

In the subtask B fine-tune setting, development and test datasets are the same as the pre-train setting. Train datasets provided contain EN and PT training examples with type of one-shot and zero-shot totally, and also contain some GL training one-shot examples. For integrity of idiom tokens, MWEs in datasets were added into the tokenizer of model. Training data have positive and negative types. The positive examples mean that sentence with an idiom (S_{MWE}) is the same as the sentence in which the idiom has been replaced by a phrase that correctly represents the meaning of the idiom in context (S_c). So their STS score are equal to 1 ($sim(S_{MWE}, S_c) = 1$). For the negative examples, a sentence with an idiom and the same sentence in which the idiom has been replaced by a phrase that incorrectly represents the meaning of the idiom in context (S_i) should have a low STS score. The score is approximately equal to the STS score between a sentence where the idiom has been replaced by a phrase that correctly represents its meaning and one wherein it incorrectly represents the meaning ($sim(S_{MWE}, S_i) = sim(S_c, S_i)$).

3.2 Evaluation Metrics

Subtask A is evaluated by the Macro F1 score between the gold labels and predictions. F1 score is defined as follows:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

Macro F1 score will calculate average of the precision and recall for all classes firstly, and does not take label imbalance into account.

The metric of subtask B is the Spearman Rank correlation to evaluate outputing STS scores, which is defined as follows:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{5}$$

$$d_i = rank(X_i) - rank(Y_i) \tag{6}$$

where n and d_i denote amount of data and difference between position of variate X and Y after sort, respectively. It computes Pearson correlation coefficient using rank of data sets. This metric uses to evaluate correlation between STS of model's output and gold label.

3.3 Implementation Details

Herein all experiments, we set seed of 4 and AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate 2e-5. Besides, batch-size was set as 32, 4, 16 in zero-shot , one-shot setting and subtask B respectively. A linear learning rate warm-up was 5% and 10% in subtask A and subtask B respectively. A max sequence length of 128 was in subtask A and pre-train setting, 512 in fine-tune setting. Each model was fine-tuned for 10 epochs, where the model exhibiting the best performance on the dev set was used to predict the test set in the competition.

For subtask A of binary classification, cross-entropy loss function was used. The methods of Figure 2 were experimented in the pre-train setting. We set our default pooling strategy was mean, and method (c) used mean pooling of last hidden layer and first-last hidden layer respectively.

In the fine-tune setting, a sentence transformer based on the pre-training model was created. The

Methods	Macro F1 score	
	dev	test
zero-shot setting		
baseline	0.6482	0.6540
mBERT(CLS of last layer)	0.6820	0.6209
mBERT(CLS of 12 layer)	0.6838	0.6030
xlm-R(CLS of last layer)	0.7129	0.6074
xlm-R(CLS of 12 layer)	0.7293	0.6369
one-shot setting		
baseline	0.8691	0.8646
mBERT(CLS)	0.8062	0.7429
xlm-R(CLS)	0.9002	0.8948

Table 1: Results of subtask A

Methods	Spearman's R (All)		Spearman's R (Idiom only)		Spearman's R (STS only)	
	dev	test	dev	test	dev	test
pre-train setting						
baseline	0.6790	0.4810	0.2187	0.2263	0.8182	0.8311
method(a)	0.6736	0.5117	0.1762	0.2582	0.8936	0.8006
method(b)	0.6484	0.5170	0.1905	0.2616	0.6991	0.6195
method(c) first-last-avg	0.7321	0.5602*	0.2111	0.2628*	0.8519	0.7990*
method(c) last-avg	0.7464	0.5650	0.2152	0.2586	0.8574	0.8044
fine-tune setting						
baseline	0.6629	0.5951	0.3459	0.3990	0.5429	0.5961
distiluse-v1	0.7514	0.5523	0.1881	0.2251	0.7939	0.7574
distiluse-v1+2losses	0.8127	0.6648	0.5097	0.4277	0.7248	0.6627
distiluse-v2+2losses	0.7882	0.6391	0.4022	0.3898	0.7154	0.6472

* *0.5602*, *0.2628*, and *0.7990* are the revised results, which are different from the results had been submitted on the evaluation. Because we found out later that there existed something wrong in our former data processing. It had be trained after correcting and used the same method and parameters as before.

Table 2: Results of subtask B

sentence transformer we chose, distiluse-base-multilingual-cased, was extended the vocabulary of MWEs and removed last dense layer was as extractor of feature. Then, mean pooling was applied to output, which generated sentence representation of 768 dimension. Besides, training data were divided into two datasets and used two loss function to fine-tune model. According to negative training data, S_{MWE}, S_c, and S_i constituted a triplet of $(anchor, pos, neg)$. Triplet loss function was used, and we set margin as 0.1 in our experiment. The positive training data only had S_{MWE} and S_c, which composed positive pairs $(anchor_i, pos_i)$. Similar to the training process of unsupervised SimCSE (Gao et al., 2021), $(anchor_i, pos_j) for (i \neq j)$ were as negative pairs. So multiple negatives ranking loss function were used in this part.

3.4 Comparative Results

All of Experimental results are listed in Tables 1 and 2. The reason for results of development and test set difference may be the inconsistent data distribution of them that GL was added in test data. However, the difference of them in zero-shot setting appeared that the sentences including of contiguous context did not make the model learn strong ability of generalization so that all the results on the test set were below the baseline. The results of one-shot setting indicated that classify the sentence and MWE by splicing them together as one texts could be integrated and obtained a better results and xlm-RoBERTa performed better.

In the results of pre-train setting, in comparative to method (a), while method (b) grasped better semantic information of idiom, it could not output good semantic similarity on STS dataset because of training method. It could be found that method (c) could ease their problem and achieved a better result compared with method (a) and (b). The results of fine-tune setting show that these two loss functions play a important role, and the model's understanding of idioms can be improved to some extent by contrastive learning. The distiluse-base-multilingual-cased also used baseline method based on $sim(S_{MWE}, S_i) = sim(S_c, S_i)$. Besides, although v2 supported more languages including GL than v1, it performed a bit poorer instead, which explains further that the model training on more languages appear loss of performance for the information that has been learned.

4 Conclusions

Herein, we discuss and experiment the methods we used in SemEval-2022 Task 2. We participated in two subtasks, idiomaticity detection and representation of idiomaticity. Each of subtasks including two settings, achieved the 19th, 6th, 5th, and 1th places in the final test sets, respectively. Our results showed that the idea introduced contrastive learning in representation of idiomaticity achieved good results, but methods of zero-shot setting in idiomaticity detection did not well and were lack of ability of generalization. We intended to explore and improve the performance of zero-shot and few-shot learning further in future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61966038. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 3551–3564.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2227–2237.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.