Cross Lingual Sentence Retrieval

Sai Sreenivas Chintha
Department of Computer Science
Indian Institute of Technology Hyderabad, Hyderabad
cs17btech11012@iith.ac.in

Ashwanth Kumar Gundeti
Department of Computer Science
Indian Institute of Technology Hyderabad, Hyderabad
cs17btech11017@iith.ac.in

ABSTRACT

We present a simple and robust method to extend existing sentence embedding models to new languages. This allows to create multilingual versions from previously monolingual models. The training is based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence. We use the original (monolingual) model to generate sentence embeddings for the source language and then train a new system on translated sentences to mimic the original model. Main advantage of this approach is that it is easy to extend existing models with relatively few samples to new languages.

KEYWORDS

Multi Lingual, Sentence Embeddings, Sentence Retrieval

ACM Reference Format:
Sai Sreenivas Chintha and Ashwanth Kumar Gundeti. 2018. Cross Lingual Sentence Retrieval. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

We present a method that allows us to extend existing sentence embeddings models to new languages. We make use of a teacher model M for source language s and a set of parallel (translated) sentences $((s_1, t_1), ...,(s_n, t_n))$ with t_i as the translation of s_i. Note that, t_i can be in different language. We train another model the student model $M^{'}$ such that $M^{'}(s_i) \equiv M(s_i)$ and $M^{'}(t_i) \equiv M(s_i)$ using mean squared loss. In this way the student model $M^{'}$ distills the knowledge of the teacher M in a multilingual setup.

The student model $M^{'}$ learns a multilingual sentence embedding space with two important properties:

- Identical sentences in different languages are close in the vector space.
- vector space properties in the original source language from the teacher model M are adopted and transferred to other languages.

This approach has various advantages compared to other training approaches for multilingual sentence embeddings. Main advantage of this approach is that it is easy and efficient because the we develop the new model is based on the already existing monolingual model.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 APPROACH

2.1 Architecture

We use a teacher model M, that maps sentences in one source language s to a vector space. And we need parallel (translated) sentences $((s_1, t_1), ...,(s_n, t_n))$ with t_i the translation of s_i. Note that, t_i can be in different languages. We train another model the student model $M^{'}$ such that $M^{'}(s_i) \equiv M(s_i)$ and $M^{'}(t_i) \equiv M(s_i)$ using mean squared loss. So our main goal is to minimize the mean-squared loss given by:

$$\frac{1}{N}\Sigma_{i \in N}[(M(s_i) - M^{'}(s_i))^2 + (M(s_i) - M^{'}(t_i))^2]$$

The architecture is illustrated in Figure 1

2.2 Dataset

In our experiments, we used the following datasets:

- **TED** It consists of 50,000 mappings for each pair of languages.
  Link for the dataset : https://sbert.net/datasets/ted2020.tsv.gz
- **BUCC** It consists of 10,000 sentences of every language and 1000 gold standard pairs .
  Link for the dataset : https://comparable.limsi.fr/bucc2017/bucc2017-task.html

The data set sizes for English-German (EN-DE), English-French (EN-AR), English-Russian (EN-RU). For training, we balance the data set sizes by picking up 50,000 parallel senteces from each dataset. We trained variations of XLM-R as our student model and used BERT fine-tuned on English NLI and STS dataset as our teacher model. We trained for a maximum of 5 epochs with batch size 128, 10,000 warm-up steps, and a learning rate of 0.00001. As development set, we measured the MSE loss and translation loss on parallel sentences. The availability of pre trained models helped in fine tuning the models with the help of small datasets.

3 EXPERIMENTS - BUCC: BITEXT RETRIEVAL

In this section, we conduct experiments on bitext retrieval i.e identifying parallel (translated) sentences from two large monolingual corpora.

We evaluate the following systems:

- bert-base-nli-stsb-mean-tokens (teacher model) :: xlm-roberta-base (student model)
- bert-base-nli-stsb-mean-tokens :: xlm-r-100langs-bert-base-nli-mean-tokens

Bitext retrieval aims to identify sentence pairs that are translations in two corpora in different languages.For our experiments,

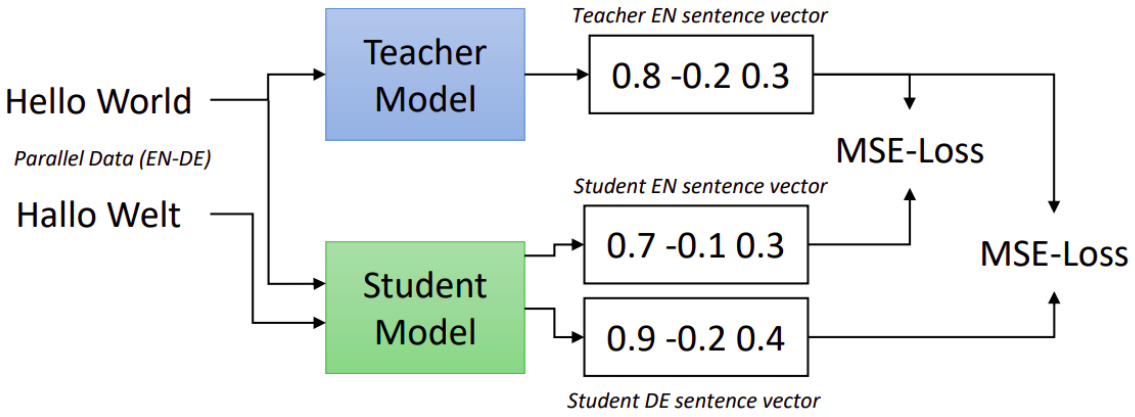

Figure 1 : A representation of the teacher-student model. Given parallel data (e.g, English German in the above figure) the student model is trained such that the produced embeddings for the English and German sentences are close to the teacher English embedding.

we use the BUCC bitext retrieval with the scoring function given by:

$$score(x, y) = margin(cosine\_sim(x, y),$$
$$\sum_{z \in NN_k(x)} \frac{cosine\_sim(x, z)}{2k} + \sum_{z \in NN_k(y)} \frac{cosine\_sim(y, z)}{2k}$$

with x, y the two sentence embeddings and $NN_k(x)$ denoting the k nearest neighbors of x in the other language. As margin function, we used $margin(a, b) = a/b$.

We used the dataset from the BUCC mining task, with the goal of extracting parallel sentences between an English corpus and four other languages: German, French, Russian, and Chinese. The data is split into training and test sets. The training set is used to find a threshold for the score function. We tried 10 different values of threshold and picked the threshold which maximizes the F1 score. Pairs above the threshold are returned as parallel sentences. Performance is measured using F1 score.

Results are shown in Table 3. Using mean pooling directly on mBERT / XLM-R produces low scores. While training on English NLI and STS data improves the performance for XLM-R (XLMR-nli-stsb), it reduces the performance for mBERT. It is unclear why mBERT mean and XLM-R mean produce vastly different scores and why training on NLI data improves the cross-lingual performance for XLM-R, while reducing the performance for mBERT. As before, we observe that mBERT / XLM-R do not have well aligned vector spaces and training only on English data is not sufficient. Using our multilingual knowledge distillation method, we were able to significantly improve the performance compared to the mBERT / XLM-R model trained only on English data.

Translation Type	Average Cosine Similarity
En-De	0.914
En-Fr	0.919
En-Ru	0.876

Table 1 : Average Cosine Similarity Scores on the TED test dataset

Models	Sim. Threshold	En-De	En-Fr	En-Ru
bert ← xlm-r-b	0.8	0.649	0.643	0.579
bert ← xlm-r-100l-b	0.75	0.853	0.862	0.748
bert ← xlm-r-100l-b	0.8	0.939	0.933	0.811

Table 2 : F1 scores on the BUCC Bitext Dataset (Score Threshold : 1.06)

4 RESULTS

After the training is completed, We have tested on test parallel datasets consisting of 2,000 pair of sentences of En-De, En-Fr, En-Ru each for calulating the cosine similarity values. The results have been reported in Table 1.

We have used the BUCC Bitext Dataset for calculating the F1 scores on the testing dataset. In the En-De dataset, there were 9,663 English sentences and 13,567 German sentences. Out of which, there were 1038 gold standard pairs. In the En-Fr dataset, there were 9,097 English sentences and 9,116 French sentences. Out of which, there were 929 gold standard pairs. In the En-Ru dataset, there were 10678 English sentences and 9927 Russian sentences. Out of which, there were 904 gold standard pairs. We have calculated the F1 scores on these datasets, using the **score** metric mentioned before. Refer to Table 2 for the values.

5 REFERENCES

1. **Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.** Nils Reimers, Iryna Gurevych

2. **Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond.** Mikel Artetxe, Holger Schwenk

3. **Pre-training via Paraphrasing.** Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, Luke Zettlemoyer