Evaluation and Compression of Different Language Transformer Models for Semantic Textual Similarity Binary Task using Minority Language Resources

Ma. Gracia Corazon Cayanan[1], Kai Yuen Cheong[2], and Li Sha[3]
graciacayanan@thei.edu.hk[1], kycheong@thei.edu.hk[2], ls050652@thei.edu.hk[3]
Faculty of Design and Environment
Technological and Higher Education Institute of Hong Kong

Abstract

Training a language model for a minority language has been a challenging task. The lack of available corpora to train and fine-tune state-of-the-art language models is still a challenge in the area of Natural Language Processing (NLP). Moreover, the need for high computational resources and bulk data limit the attainment of this task. In this paper, we presented the following contributions: (1) we introduce and used a translation pair set of Tagalog and English (TL-EN) in pre-training a language model to a minority language resource; (2) we fine-tuned and evaluated top ranking and pre-trained semantic textual similarity binary task (STSB) models, to both TL-EN and STS dataset pairs. (3) then, we reduced the size of the model to offset the need for high computational resources. Based on our results, the models that were pre-trained to translation pairs and STS pairs can perform well for STSB task. Also, having it reduced to a smaller dimension has no negative effect on the performance but rather has a notable increase on the similarity scores. Moreover, models that were pre-trained to a similar dataset have tremendous effect on the model's performance scores.

Keywords: semantic matching, semantic textual similarity binary task, low resource minority language, fine tuning, dimension reduction, transformer models

1. Introduction

Semantic Matching is an area in Natural Language Processing that deals with the estimation between a source and a target (Jiang, et.al, 2019). The state-of-the-art evolution of this area was evidently shown in research publications, from simple text matching (Liu, et.al. 2016), short text matching (Rao, et.al. 2019), enhanced text matching transformation (Zhang, et.al. 2019), heuristic matching (Wu, et.al. 2017), to long-form document matching (Jiang et.al. 2019). The contributions of these papers has built a good reputation and pioneered different areas to challenge other experts of the field. Due to its major contribution to the overall success of information retrieval tasks in NLP, challenges became evident and as Jiang, et.al. (2019) emphasized, semantic matching is the 'holy grail' in textual information retrieval.

One of the challenging tasks under semantic matching is the semantic textual similarity binary task (STSB). This task makes use of sentence pairs with binary similarity labels. Language models are mostly trained in English language sentence pairs for STSB. This is because of the profound collection of the available datasets in the English language. Minority languages such as Tagalog and Cantonese has limited to no record of training language models for a semantic text matching task (Singh and Klakow, 2013), specifically, STSB. This is due to its limited dataset collection to no collection at all.

In training language models for STSB task whether with the use of large collection or low-resource dataset, both requires high computational resource. This is to accommodate longer inference time, higher memory space and higher processing units. Techniques such as hyper parameter tuning (Cruz and Cheng, 2019; Yang and Shami, 2020), dimension reduction (Mehta, et al., 2019; Heo, et.al., 2020), and knowledge distillation (Pan and Yang, 2009; Tan, et.al, 2019; Gour, et.al, 2019) are being used to offset these demand of resources.

This study aims to train and fine-tune language models using a minority language and determine the performance of these models in a low-resource data set. Specifically, the contribution of this paper are the following:

(1) Introduce a Tagalog-English corpus that will be used in training the models to the Tagalog language, to contribute to the scarcity of low-resource dataset of minority languages;
(2) Train and evaluate language transformer models in a low-resource minority language for semantics textual similarity binary task; and
(3) Evaluate the language transformer models after a dimension reduction, offsetting the demand of high resources.

2. Methods

This part of the paper presents the requirements and techniques used. A brief discussion is provided while a detailed process of how these methods are used is discussed in the latter part of the paper.

2.1. Datasets

The study made use of two (2) different types of paired data sets. First is a translation dataset composed of English and Tagalog pairs of sentences while the second one is the sentence pairs with similarity labels of 0s and 1s. 0 indicating no similarity while 1 indicating similarity. These two (2) datasets were used to train and evaluate the transformer models also using various compression techniques. Specifically, the datasets were derived and used as follows:

English – Tagalog (EN-TL) Pairs Dataset. The EN-TL pairs dataset is a set of sentences in Tagalog and English. The Tagalog sentences was scraped from news articles from different Philippine news sites online. It is then translated to English through the Google translate package in python. The EN-TL dataset is composed of 500,000 pairs, considered to be the largest English-Tagalog dataset pairs available in NLP.

Another dataset that was used in this study is the ***STS Pairs dataset***, provided by Cruz et.al. (2019). The NewsNLI Ph was collated from the Wikipedia and transformed by the original authors into sentence pairs where similarity labels are provided. This dataset is the largest corpus of sentence similarity pairs that is made available online. The study took the privilege of using it to limit the use of resources in making a new one.

The dataset are split into three (3) chunks, 1k pairs, 5k pairs and 10k pairs for both the translation and STSB pairs. Table 1 shows the summary of the dataset in words and in sentences.

Table 1. Summary of the Dataset

Translation Pairs (TL-EN)			
	1k-1k	5k-5k	10k-10k
words	55,249	267,971	548,905
sentences	2,939	14,101	28,164
STSB Pairs			
	1k-1k	5k-5k	10k-10k
words	74,718	283,664	564,416
sentences	2,816	14,211	28,350

2.2. Models

Transformer models are language models Vaswani et.al. (2017) specifically designed for pre-training the bidirectional structure of encoder-decoder and can be fine-tuned to several NLP tasks.

Transformer models uses 'Attention' as a mechanism in prioritizing certain tokens in a sequence by giving more weights to more essential parts. As Vaswani et al (2017) describes it, "Attention is paying more attention to more important parts".

Pre-trained transformer models were made available in public through the Hugging face repository online. Hugging face (https://huggingface.co/) is an open community of NLP experts where ideas, outputs and results are shared. Pre-trained models in different NLP tasks are available in this repository. One of the tasks that these models that are pre-trained to is the Semantic Textual Similarity Binary Task (STSB). STSB is a semantic matching task that checks the similarity of sentence pairs and uses binary similarity labels, 0s and 1s. A list of high-performing transformer models was provided by Hugging face. As shown in figure 1, the top three (3) models that were trained to STSB tasks are:

(1) stsb-mpnet-base-v2,
(2) stsb-roberta-base-v2, and
(3) stsb-distilroberta-base-v2

This paper evaluated these top three (3) models in the STSB task using the Tagalog dataset. The paragraphs below describe each of these models and how each of them become state-of-the-art models placing on top of other language models evaluated in the STSB task.

Model Name	Base Model	Pooling	Training Data	STSb Performance (Higher = Better)
stsb-mpnet-base-v2	mpnet-base	Mean Pooling	NLI+STSb	88.57
stsb-roberta-base-v2	roberta-base	Mean Pooling	NLI+STSb	87.21
stsb-distilroberta-base-v2	distilroberta-base	Mean Pooling	NLI+STSb	86.41
nli-mpnet-base-v2	mpnet-base	Mean Pooling	NLI	86.53
stsb-roberta-large	roberta-base	Mean Pooling	NLI+STSb	86.39
nli-roberta-base-v2	roberta-base	Mean Pooling	NLI	85.54
stsb-roberta-base	roberta-base	Mean Pooling	NLI+STSb	85.44
stsb-bert-large	bert-large-uncased	Mean Pooling	NLI+STSb	85.29
stsb-distilbert-base	distilbert-base-uncased	Mean Pooling	NLI+STSb	85.16
stsb-bert-base	bert-base-uncased	Mean Pooling	NLI+STSb	85.14
nli-distilroberta-base-v2	distilroberta-base	Mean Pooling	NLI	84.38

Figure 1. Transformer Models pre-trained in STSB Task

(1) stsb-mpnet-base-v2 was derived from its base model, Microsoft's MPNet, Masked and Permuted Pre-training for Language Understanding (Song, et.al., 2020). It provides better accuracy through combining the power of masked and permuted language modeling. Table 2 shows that the MLM (masked language modeling) models predict the tokens [sentence, classification] independently which may incorrectly predict the classification token.

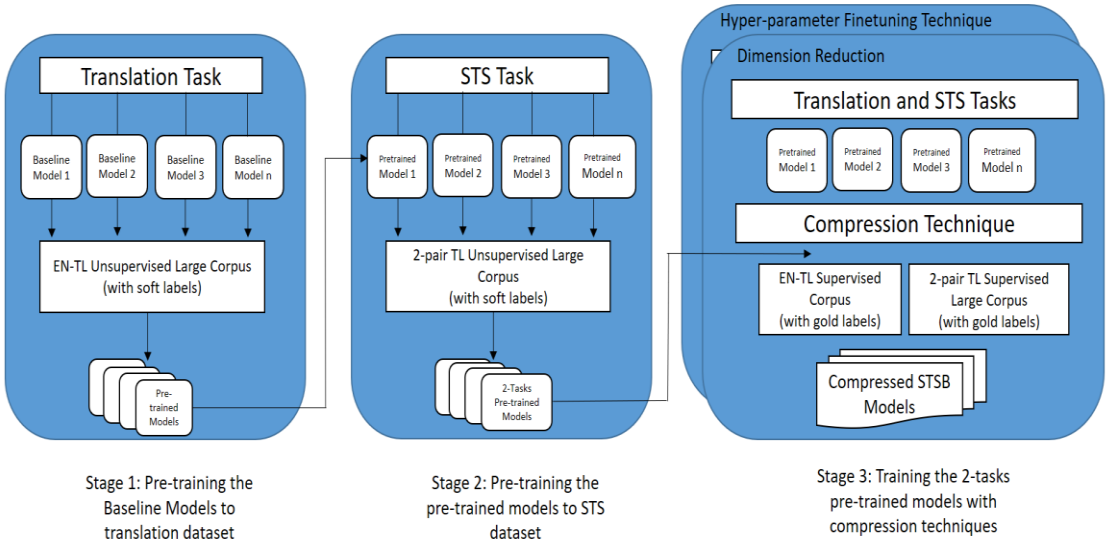

Figure 2. Architecture of Training Multi-models to Minority Language Resource with Model Compression Techniques

PLM (permuted language modeling) predicts also [sentence, classification] without the position information of the whole sentence, thus, resulting into incorrectly predicting three (3) tokens, [sentence, pair, classification]. MPNet can predict using both the full position information, and the dependence of the tokens to each other, thus providing a more accurate prediction.

Table 2. Difference in factorization of MLM, PLM and MPNet (Song et.al. 2020)

Type	Factorization
MLM (Masked)	*log P* (sentence | the task is [M] [M]) + *log P* (classification | the task is [M] [M])
PLM (Permuted)	*log P* (sentence | the task is) + *log P* (classification | the task is sentence)
MPNet (Masked and Permuted)	*log P* (sentence | the task is [M] [M] + *log P* (classification | the task is sentence [M])

stsb-mpnet-base-v2 is a pre-trained model that maps sentences and paragraphs to a 68 dimensional dense vector space. The model has a size of 418 MB, max sequence length of 75, and uses mean pooling.

(2) stsb-roberta-base-v2 was derived from its base model, RoBERTa, Robustly Optimized BERT Pre-training Approach (Liu, et.al., 2019). This model was pre-trained using the base model BERT. RoBERTa was fine-tuned and trained using a total of 160 GB of text. One of which is a novel dataset, called CC-News, containing 63 million English news articles. The model mainly used the MLM technique and next sentence prediction. The lack of training of BERT was solved by RoBERTA where training techniques include dynamic masking, full-sentences without NSP (Next Sentence Prediction) losses, large mini-batches and a larger byte-level BPE (Byte-Pair Encoding). This model is case sensitive, making 'Word' and 'word' different.

stsb-roberta-base-v2 is a pre-trained model that also maps sentences in a 768 dimensional vector space. It has a max length of 75, and also uses mean pooling.

(3) stsb-distilroberta-base-v2, was derived from a distilled version of RoBERTa, DistilRoBERTa, which has 6 layers, 768 dimension and 12 heads, with a total of only 82M parameters (as compared to the original base model of 125M). This makes DistilRoBERTa twice as fast as its base model. DistilRoBERTa was pre-trained on an OpenWebCorpus (Gokaslan and Cohen, 2019), an open source dataset from the original OpenAI's WebText dataset. It was extracted from all Reddit post urls, containing web page contents. This dataset is four times less from the training data of RoBERTa.

3. Experiments and Results

Figure 2 presents the experimental setting of this paper where there are three (3) stages:

(1) Pre-training the selected models using the translation dataset
(2) Then, pre-train these trained from the translation datasets to STS dataset.
(3) Lastly, conduct compression techniques and evaluate the performance of the models after compression.

This section details the experimental setup and the results from each step.

Table 3. Fine-tuning Results

Model Type	Data Splits	MSE Loss	Pearson's r	Spearman's p
stsb-mpnet-base-v2	1k-1k	**0.4027**	0.3021	0.3385
stsb-roberta-base-v2	1k-1k	0.6618	**0.4326**	**0.4601**
stsb-distilroberta-base-v2	1k-1k	0.5091	0.3073	0.3689
stsb-mpnet-base-v2	5k-5k	0.9257	0.1151	0.1999
stsb-roberta-base-v2	5k-5k	0.9271	**0.39021**	**0.4449**
stsb-distilroberta-base-v2	5k-5k	**0.7715**	0.1965	0.3486
stsb-mpnet-base-v2	10k-10k	0.0926	0.1151	0.1999
stsb-roberta-base-v2	10k-10k	**0.1122**	**0.3844**	**0.4294**
stsb-distilroberta-base-v2	10k-10k	0.8805	0.2453	0.3274

Table 4. Cosine Similarity Results after Model Reduction to 128 Dimensions

Model Type	Data Splits	Pearson's r	Spearman's p
stsb-mpnet-base-v2	1k-1k	0.3017	0.3412
stsb-roberta-base-v2	1k-1k	**0.4415**	**0.4688**
stsb-distilroberta-base-v2	1k-1k	0.2771	0.2964
stsb-mpnet-base-v2	5k-5k	0.1535	0.2475
stsb-roberta-base-v2	5k-5k	**0.3931**	**0.4491**
stsb-distilroberta-base-v2	5k-5k	0.1960	0.3565
stsb-mpnet-base-v2	10k-10k	0.1126	0.2022
stsb-roberta-base-v2	10k-10k	**0.3899**	**0.4364**
stsb-distilroberta-base-v2	10k-10k	0.2503	0.3366

3.1. Pre-training the Models

Models were pre-trained in a large corpora, unsupervised corpus of 500,000 rows of Tagalog-English translation sentence pairs and 100,000 rows STS pairs with binary soft labels using the NewsNLI dataset (Cruz and Cheng, 2019). Both corpora was derived from news articles of Philippine websites.

The models were pre-trained with a learning rate of 5e-4, batch size of 256, warmup steps of 1000 and evaluation steps of 1000. The models were pre-trained in 20 epochs. The result of the pre-training is as follows:

Table 5. Pre-training Results

Model	MSE Loss	Pearson's r	Spearman's p
stsb-mpnet-base-v2	**0.0199**	0.3127	0.3248
stsb-roberta-base-v2	0.1432	**0.4239**	**0.4274**
stsb-distilroberta-base-v2	0.1132	0.2929	0.3082

As shown in the Table 5 above, the stsb-mpnet-base-v2 garnered the lowest score in the Mean Squared Error (MSE) indicating that the data loss is minimal as compared to the other two (2) models. Moreover, in the initial evaluation of the models after the training, stsb-roberta-base-v2 garnered the highest score in the Cosine Similarity Scores both in Pearson's r and Spearman's p.

Interesting observation in the result is, despite the highest loss value of stsb-roberta-base-v2 as compared to the other two (2) models, the result in its performance for STSB task is higher on both scores of the cosine similarity. While stsb-mpnet-base-v2 has the lowest loss score and supposed to be the highest performer in the field of STSB, it ranked second on both similarity scores.

3.2. Fine-tuning the Models

After the pre-training, the models were fine-tuned to smaller versions of both the translation pairs and STS pairs corpora. This is to examine the performance of the models in a small collection of datasets. The fine-tuning was conducted to different data splits of 1k, 5k and 10k. The fine-tuning was done independently to each dataset and to each model.

Fine-tuning ran in five (5) epochs, warm up and evaluation steps at 1000 and with optimization values of learning rate: 2e-5 and epsilon: 1e-6.

Based on the results shown in table 3, model *stasb-roberta-base-v2* outperformed the two other models from all splits of data sets on both cosine similarity scores of Pearson's r and Spearman's p.

Moreover, the result of the MSE loss is inconclusive as it shows different models with high number of losses during fine-tuning.

Table 6. Overall Performance of stsb-roberta-base-v2

Model Type	Data Splits	Pearson's r	r Decrease	Spearman's p	p Decrease
stsb-roberta-base-v2 fine-tuned	1k-1k	0.4326	-	0.4601	-
	5k-5k	0.39021	0.0424	0.4449	0.0152
	10k-10k	0.3844	0.0058	0.4294	0.0155
stsb-roberta-base-v2 128 dimension	1k-1k	0.4415	-	0.4688	-
	5k-5k	0.3931	0.0484	0.4491	0.0197
	10k-10k	0.3899	0.0032	0.4364	0.0127

Table 7. Increase in Performance from Fine-tuning to Dimension Reduction

Cosine Similarity Score	Data Splits	Fine-tuned	128 Dimension	Score Increase
Pearson's r	1k-1k	0.4326	0.4415	+0.0089
	5k-5k	0.3902	0.3931	+0.0029
	10k-10k	0.3844	0.3899	+0.0055
Spearman's p	1k-1k	0.4601	0.4688	+0.0087
	5k-5k	0.4449	0.4491	+0.0042
	10k-10k	0.4294	0.4364	+0.007

3.3. Dimension Reduction

After the fine-tuning, the model is reduced to 128 dimension. This is to check how the model performs even when it is compressed from 768 to 128 dimension. The objective of this is to offset the need for higher requirements in running state-of-the-art language models.

Based on the results shown in Table 4, stsb-roberta-base-v2 again consistently outperformed the two other models in all dataset splits, even after the dimension size was reduced.

3.4. Overall Result and Insights

The results show that the top three (3) transformer language models can perform well in a semantic textual similarity binary classification task using a low-resource minority language such as the Tagalog dialect, one of the most spoken dialects in the Philippines. In the original ranking of the models, it was stsb-mpnet-base-v2 that ranked above all models as shown in Figure 1. But in this paper, stsb-roberta-base-v2 outperformed stsb-mpnet-base-v2 from the results of the pre-training to the results of all the experiments conducted. Recalling the datasets that stsb-roberta-base-v2 was trained to, CC-News is a collection of 63 million news articles. And, the dataset that this paper used also comes from news articles that was scrapped from different Philippine news websites. The similarity of these datasets, used to pre-train the model and the content of the dataset that this paper used, had provided a major impact to the results and with that, outperforming other models that were not trained to the same dataset as stsb-roberta-base-v2 had.

Moreover, looking at the results of stsb-roberta-base-v2, shown in Table 6, a decrease of the cosine similarity scores were seen when there is an increase of the data splits. This is both seen in the fine-tuning and in the dimension reduction.

However, stsb-roberta-base-v2 has an increase of its score from fine-tuned to reduction of its dimension to 128 (Table 7). This proves that even after reducing its size, stsb-roberta-base-v2 can still perform well for STSB task in low-resource minority language.

4. Conclusion

This paper shows that language models that are pre-trained, fine-tuned and compressed to a smaller dimension can be used in a low-resource minority language, specifically, for the semantic text similarity binary classification task. This is through the use of two (2) kinds of datasets: the translation pairs and the STS pairs.

Moreover, the size of the model being reduced to a smaller dimension does not have a direct effect on its performance as long as it was pre-trained and fine-tuned prior the reduction. With this, models of smaller dimension can reduce the use of high and expensive computational resources.

However, models that top ranked in a specific task may not consistently outperform other models in other datasets. This paper shows that the similarity of the datasets impacted the scores of the models. A model with the same pre-trained dataset can

outperform other models which were pre-trained to a different type of dataset. In the case of using a dataset with news article contents, stsb-roberta-base-v2 performs well due to its pre-training to a large corpus called CC-News.

6. Acknowledgment

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. UGC/FDS25/E07/19). Any opinions, findings, and conclusions of this paper are those of the authors and do not necessarily reflect the views of the sponsor.

References:

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. MPNet: Masked and Permuted Pre-training Language Understanding. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. arXiv: 2004.09297v2 [cs.CL]. 2 Nov 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veseliin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Paul G. Alen School of Computer Science and Engineering. arXiv:1907.11692v1 [cs.CL]. 26 Jul 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All You Need. arXiv:1706.03762v5 [cs.CL] 6 Dec 2017.

Nils Reimers and Iryna Guvervych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Arxiv:1908.10084v1 [cs.CL]. 27Aug 2019.

Nandan Thakur, Nils Reimers, Johannes Daxenberger and Iryna Gurevych. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. arXiv:2010.08240v2 [cs.CL]. 12 Apr 2021.

Jan Christian Blaise Cruz and Charibeth Cheng. Evaluating Language Model Finetuning Techniques for Low-resource Languages. arXiv:1907.00409v1 [cs.CL] 30 Jun 2019

Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco and Charibeth Cheng. Exploiting News Article Structure for Automatic Corpus Generation. arXiv:2010.11574v2 [cs.CL] 20 May 2021.

Pengfei Liu, Xipen Qiu, Jifan Chen and Xuanjing Huang. Deep Fusion LSTMs for Text Semantic Matching. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1034-1043, Berlin, Germany. 7-12 August 2016.

Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi and Jimmy Lin. Bridging the Gap Between Relevance Matching and Semantic Matching for Short Text Similarity Modeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5370-5381. Hong Kong, China. 3-7 November 2019.

Shutao Zhang, Haibo Tan, Liangfeng Chen, and Bo LV. Enhanced Text Matching Based on Semantic Transformation. IEEE Access.2020.2973206. 19 February 2020.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, Mark Najork. Semantic Text Matching for Long-Form Documents. ACM ISBN978-1-4503-6674-8/19/05. DOI: 101145/3308558.3313707. 19 May 2019.

Zongda Wu, Hui Zhu, Guiling Li, Zongmin Cui, Hui Huang, Jun Li, Enhong Chen, Guandong Xu. An Efficiet Wikipedia Semantic Matching Approach to Text Document Classification. Elsevier: Information Systems DOI:S00200255. 7 February 2017.

Byeongho Heo, Sangdoo Yun, Dongyoon Hun, Sanghyuk Chun, Junsuk Choe, Seong Joon Oh. Rethinking Spatial Dimensions of Vision Transformers. Computer Vision Foundation Open Access. IEEE Xplore.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer and Hannaneh Hajishirzi. DeLight: Deep and Light-Weight Transformer. ICLRI 2021.

Li Yang and Abdallah Shami. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. arXiv:2007.15745v2 [cs.LG]. 7 Aug 2020.

Mittul Singh, Dietrich Klakow. Comparing RNNs and Log-Linear Interpolation of Improved Skip-Model on Four Babel Languages: Cantonese, Pashto, Tagalog, Turkish. DOI:978-1-4799-0986 @2013 IEEE.

Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp. A Multilingual Dataset for Evaluating Parallel Sentence Extraction from Comparable Corpora.

Hua He, Kevin Gimpel, and Jimmy Lin. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks.

S. Gutstein, O. Fuentes, and E. Freudenthal, “Knowledge transfer in deep convolutional neural nets,” International Journal on Artificial Intelligence Tools, vol. 17, no. 03, pp. 555–567, 2008.

C. Bucilua, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 535–541.

G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” NIPS Deep Learning Workshop, 2015