JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Advancing Scalabale Attentive Sentence-Pair Modeling with "RoBERTa-GPT"

Hutesh Vidhate,
Dept.of computer engineering,
COEP technological University,
Pune,India

ABSTRACT

In this paper, we introduce a novel framework for enhancing scalable attentive sentence-pair modeling by synergistically integrating RoBERTa and GPT architectures. RoBERTa, known for its robust pretraining methodology and fine-tuned representations, is merged with GPT's generative capabilities to create a unified model that effectively captures contextual understanding and fosters generative fluency. Through extensive experimentation on benchmark datasets, our proposed RoBERTa-GPT fusion framework demonstrates superior performance and scalability across various sentence-pair tasks, showcasing its potential for advancing the state-of-the-art in natural language processing.

INTRODUCTION

The comprehension and interpretation of textual data lie at the core of numerous natural language processing (NLP) tasks, including sentiment analysis, question answering, and machine translation. Among these, attentive sentence-pair modeling emerges as a pivotal area, facilitating tasks such as paraphrase identification, textual entailment, and semantic similarity assessment. The essence of attentive sentence-pair modeling lies in its ability to discern nuanced relationships between pairs of sentences, enabling machines to comprehend semantic similarities and differences effectively. Traditional approaches to sentence-pair modeling often relied on handcrafted features and shallow learning algorithms, which struggled to capture the complexity and variability inherent in natural language. However, recent advancements in deep learning, particularly transformer-based architectures, have revolutionized the landscape of sentence-pair modeling. Models such as RoBERTa and GPT have demonstrated exceptional prowess in capturing contextual information and generating coherent text, respectively, paving the way for innovative approaches to attentive sentence-pair modeling. In this paper, we embark on a journey to advance the frontier of attentive sentence-pair modeling by synergistically integrating the robust pretraining methodology of RoBERTa with the generative capabilities of GPT. Our proposed framework, coined "RoBERTa-GPT Fusion," aims to harness the complementary strengths of these architectures to create a unified model that excels in capturing contextual understanding and fostering generative fluency in sentence-pair tasks**.** Recently, there have been various neural net- work models proposed for sentence pair modelingtasks, including semantic similarity (Agirre et al., 2015), paraphrase identification (Dolan et al.,2004; Xu et al., 2015), natural language infer-ence (Bowman et al., 2015), etc. Most, if not all, of these state-of-the-art neural models (Yin et al., 2016; Parikh et al., 2016; He and Lin,2016; Tomar et al., 2017; Shen et al., 2017) have achieved the best performances for these tasks by using pretrained word embeddings, but re-sults without pretraining are less frequently re- ported or noted. In fact, we will show that, even with fixed randomized word vectors, the pairwiseword interaction model (He and Lin, 2016) based on contextual word vector similarities can still achieve strong performance by capturing identi- cal words and similar surface context features. Moreover, pretrained word embeddings generally have poor coverage in social media domain whereout-of-vocabulary rate often reaches over 20% (Baldwin et al., 2013).

We investigated the effectiveness of sub- word units, such as characters and character n-grams, in place of words for vector repre- sentations in sentence pair modeling. Though it is well-known that subword representa- tions are effective to model out-of-vocabularywords in many NLP tasks with a single sentence input, such as machine translation (Luong et al., 2015; Costa-jussa` and Fonollosa, 2016), language modeling (Ling et al., 2015; Vania and Lopez, 2017), and sequence labeling (dos Santos and Guimarães, 2015; Plank et al., 2016), they are not systematically studied in the tasks that concern pairs of sentences. Un- like in modeling individual sentences, subword representations have impacts not only on the out-of-vocabulary words but also more directly on the relation between two sentences, which is calculated based on vector similarities in many sentence pair modeling approaches (more details in Section 2.1). For example, while subwords may capture useful string similarities between a pair of sentences (e.g. spelling or morphological variations: *sister* and *sista*, *teach* and *teaches*), they could introduce errors (e.g. similarly spelled words with completely different meanings: *ware* and *war*).

1. Related Work

There have recently appeared an increasing number of stud- ies suggesting usage of general language representation models for natural language understanding tasks. Among the most promising techniques, the unsupervised fine-tuning approach has been shown to be effective on many sentence-level tasks (Dai and Le 2015; Howard and Ruder 2018; Rad- ford et al. 2018). This technique uses a sentence encoder to produce contextual token representations. The encoder training procedure is composed of two phases: (1) unsuper- vised training on unlabeled text, and (2) fine-tuning for su- pervised downstream tasks. The unsupervised training al- lows the model to learn most of the parameters in advance, leaving only few parameters to be learned from scratch dur- ing fine-tuning.

More recently, BERT (Devlin et al. 2019) has emerged asa powerful method that has achieved state-of-the-art resultsin various sentence or sentence-pair language understanding tasks from the GLUE benchmark (Wang et al. 2018), includ- ing sentiment analysis (Socher et al. 2013), paraphrase iden- tification (Williams et al. 2017) and semantic text similarity (Cer et al. 2017). Liu et al. (Liu et al. 2019), introduce Multi-Task Deep Neural Network (MT-DNN), which extends BERT by learning text representations across multiple nat- ural language understanding tasks. In sentence-pair tasks, both BERT and MT-DNN require feeding both sentences together as a single input sequence. While other techniques, such as (Conneau et al. 2017; Subramanian et al. 2018), sug- gest extracting a feature vector for each sentence separately via an embedding function, followed by a relatively low cost similarity function which produces a similarity score for the vector-pair.

2. Distilled Sentence Embedding (DSE)

In this section, we present the problem setup and describe the DSE model in detail.

3.1 Problem Setup

Let $\mathrm{G} = \{w_i\}_{i=1}^{w}$ be the vocabulary of all supported tokens. We define Y to be the set of all possible sentences that can be generated using the vocabulary $\mathbf{G}$.

Let T: $Y \times Y \rightarrow \mathbb{R}$ be the teacher model (e.g., a fine-tuned BERT model). T receives a sentence-pair $(y, z) \in Y \times Y$ and outputs a similarity score $T_{yz} \triangleq T(y, z)$. Note that T is not necessarily a symmetric function.

Let ψ, ϕ: $Y \rightarrow \mathbb{R}^d$ be sentence embedding functions that embed a sentence $y \in Y$ in a d-dimensional latent vector space. The usage of different sentence embedding func- tions, ψ and ϕ, is due to the fact that T is not necessarily a symmetric function. For example, in BERT, the sentences A and B are associated with different segment embeddings. Therefore, ψ and ϕ play a similar role as the common con- text and target representations that appear in many neural embedding methods (Barkan 2017; Barkan and Koenigstein2016; Mikolov et al. 2013; Mnih and Hinton 2009).

Let f: $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a (parametric) similarity func- tion. f scores the similarity between sentence embeddings that are produced by ψ and ϕ. Then, the student model

S: $Y \times Y \rightarrow \mathbb{R}$ is defined as

$$S_{yz} \triangleq f(\psi(y), \phi(z)). \qquad (1)$$

Pairwise Training

In pairwise training, we define a loss function f: $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ and train S to minimize $f(S_{yz}, T_{yz})$ in an end-to-end fash-ion. Specifically, given a sentence-pair $(y, z) \in X \times X$, we compute the embeddings $\psi(y)$ and $\phi(z)$ for the sentences $y$and z, respectively. Then, the similarity score S_{yz} is com- puted using the similarity function f according to Eq. (1).

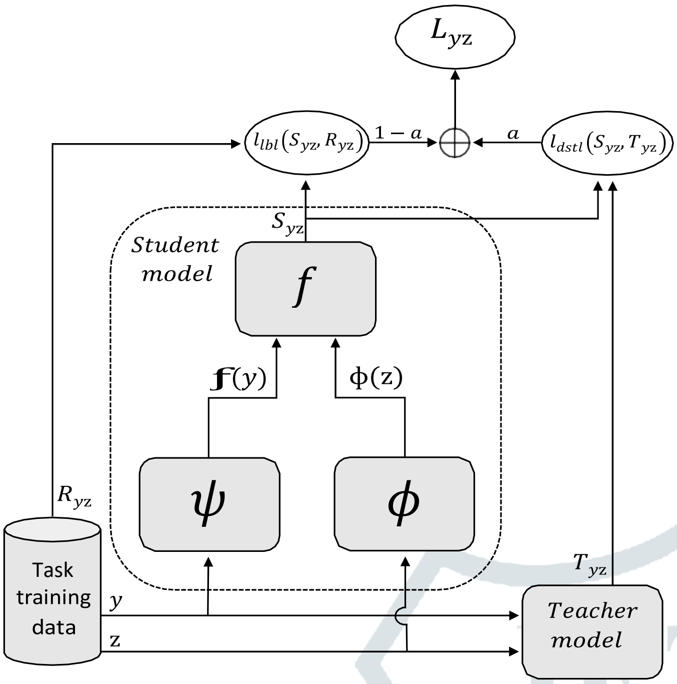
Note that f can be either a regression or classification loss depending on the task at hand. Moreover, f can be trivially extended to support multiple teacher models. In (Hinton et al. 2014) the authors suggest using two teacher models T and R, where R is simply the ground truth labels as follows

$$f_{yz} = \alpha l_{dstl}(S_{yz}, T_{yz}) + (1-\alpha) l_{lbl}(S_{yz}, R_{yz}) \quad (2)$$

where $\alpha \in [0,1]$ is a hyperparameter that controls the rela- tive amount of supervision that is induced by T and R. In this case, the student model is simultaneously supervised by

T and R. Note that in general, the distillation loss l_{dstl} and the ground truth label loss l_{lbl} are not restricted to be the same loss function (as shown in Section 3.5). The DSE model is illustrated in Fig. 1.

(2)



3.1 The Teacher Model

The teacher model T is implemented as a BERT-Large model from (Devlin et al. 2019), consisting of 24 encoder layers that each employ a self-attention mechanism. For a sentence-pair input, T employs CA between the two sen- tences. The teacher model is initialized to the pre-trained version from (Devlin et al. 2019) and then fine-tuned ac- cording to each specific sentence-pair task.

After the fine-tuning phase, we compute the score T_{yz} fora sentence-pair (y, z) by propagating a unified representa- tion of the sentence-pair throughout T, as done in (Devlin etal. 2019). The score is then extracted from the output layer,which is placed on top of the last hidden representation of the CLS token. Note that T_{yz} is set to the logit value (before the softmax / sigmoid activation).

It is important to emphasize that DSE is not limited to BERT as a teacher model. For example, we could use the exact same method with an XLNet (Yang et al. 2019)teacher. The choice of BERT is mainly due to its prevalence

Experimental Setup and Results

We evaluate DSE in two different settings: First, task specific embeddings for sentence-pair tasks, where the whole model is trained in an end-to-end fashion and evaluated on a specific dataset. Second, universal sentence representa- tions generation, in which the model is pre-trained to pro- duce general purpose sentence embeddings. In addition, wereport empirical results that showcase the efficiency of DSE in computing sentence-pair similarities compared to ROBERT-GPT.

4.1 Sentence-Pair Modeling

For sentence-pair tasks, our evaluation includes several datasets from the GLUE benchmark: MRPC (Dolan and Brockett, 2005), MNLI (Williams et al., 2018), QQP, QNLI (Wang et al., 2018), and STS-B (Cer et al., 2017). These datasets represent different tasks that revolve around modeling and scoring sentence-pairs. MRPC, STS-B, and QQP focus on semantic similarity of phrases or questions, MNLI is a natural language inference (NLI) benchmark, and lastly, QNLI is a question answering dataset. We refer to (Wang et al. 2018) for a detailed description of these datasets.

BERT-Large: This is the BERT-Large model from (Devlin et al. 2019). This model is also used as a teacher model. Results are reported from (Devlin et al. 2019).

BERT-Base: This is the BERT-Base model from (Devlinet al. 2019). Results are reported from (Devlin et al. 2019).

DSE: This is our proposed model from Section 3. We consider three variants of DSE that differ by the parameter values of $\alpha \in \{0,0.5,1\}$ which controls the amount of distillation. For all datasets we set the distillation loss $l_{dstl} = l_{L2}$.For QQP, MRPC, QNLI and MNLI we set the label loss $l_{lbl} = l_{cce}$. Specifically, for MNLI we further used $w \in \mathbb{R}^{3\times512}$ in Eq. (3) to support a 3-dimensional output. For STS-B, we set $l_{lbl} = l_{L2}$. We used the Adam optimizer (Kingma and Ba 2014) with minibatch size of 32 and a learning rate of 2e-5, except for STS-B, where we used a learning rate of 1e-5. The models were trained for 8 epochs.The best model was selected based on the dev set.

DSE (Frozen $\boldsymbol{f}$): We trained another version of DSE in which ψ is frozen. Since ψ is implemented as BERT (Sec- tion 3.4), we further want to investigate the actual benefit from fine-tuning ψ w.r.t. the task at hand. Therefore, we pre- sent results for a DSE version in which ψ is not fine-tuned. Note that the parametric similarity function is still learned in this version.

ELMO + Attn: This is the BiLSTM + ELMO, Attn model from (Wang et al. 2018). It comes in two variants: Single-Task (ST) and Multi-Task (MT) Training. The re- sults are reported taken from (Wang et al. 2018).

GenSen: Since DSE is a sentence embedding model, we further compare its performance with GenSen (Subramanian et al. 2018), which is the best performing sentence embed- ding model from (Wang et al. 2018). The results are taken from (Wang et al. 2018).

4.1.1 Sentence-Pair Tasks Results

Table 1 presents the results for each combination of model and dataset. In addition, we provide the average score that is computed across the datasets for each model (AVG column). The last two columns present the relative degradation compared to BERT-Large and the relative improvement obtained by DSE ($\alpha = 0.5$) over each model (reported in percentages).

First, we compare between the four DSE variants. We seethat for MNLI, QNLI, MRPC and QQP, enabling distillation ($\alpha \in \{0.5,1\}$) slightly improves upon using $\alpha = 0$. How- ever, on STS-B, distillation seems to hurt performance. Weattribute the degradation to the fact that STS-B is a regres- sion task and therefore the ground truth labels are already provided in a resolution that is finer than binary values. Lastly, we see that the frozen version of DSE performs much worse than all other DSE variants. This is evidence for the importance of fine-tuning ψ, which further confirms that a naïve use of pretrained BERT for sentence embedding pro-duces relatively poor results, in some cases. Therefore, we conclude that the distilled version of DSE ($\alpha \in \{0.5,1\}$) per-forms the best. From now on, we focus on a comparison be-tween the $\alpha = 0.5$ version of DSE and the other models. Next, we turn to consider the performance gaps between DSE and BERT. Recall that DSE is supervised by BERT- Large and hence the performance gaps between the two models quantifies the ability of the former to reconstruct thelatter's scores. We see that the largest and smallest relative degradations occur on the MNLI and STS-B datasets, re- spectively. Overall, DSE results in an average relative deg-radations of 4.6% and 3.1% compared to BERT-Large and BERT-Base, respectively.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS-B	AVG
GenSen	82.5	87.7	94.0	90.9	83.2	93.0	84.4/78.6	0.888	87.8	78.9/78.6	86.8
InferSent	81.1	86.3	92.4	90.2	84.6	88.2	83.1/76.2	0.884	86.3	75.8/75.3	85.2
BERT-Large	83.5	88.8	95.5	89.1	87.1	93.2	83.5/76.4	0.838	82.2	68.4/68.3	85.1
DSE ($\alpha = 0.5$)	83.6	90.2	93.6	89.8	91.0	91.8	83.8/77.9	0.856	86.7	70.7/71.4	86.4
DSE ($\alpha = 0$)	83.1	89.8	93.1	89.4	88.3	92.0	81.8/76.2	0.847	86.1	73.1/74.1	85.9

Table 2: Universal sentence embedding benchmarks results. The evaluation results are of linear models trained over each of the model's sentence representations. The results for Gensen and InferSent are taken from their respective papers. We report the F1/accuracy scores for MRPC, Pearson correlation for SICK-R, Pearson/Spearman correlations for STSB, and accuracy for the rest. AVG column presentsthe average score across all datasets, where each dataset's score is the mean of its one or two reported scores

4.1.2 Downstream Tasks Results

For each sentence embedding method and dataset included in the evaluation, Table 2 contains the results of a shallow linear model trained on top of the precomputed embed- dings. We report results for our approach with $\alpha = 0.5$, which showed the most promising performance in Section 4.1.2, and compare it to the current state-of-the-art methods: Infersent (Conneau et al. 2017) and Gensen. Additionally, we include a comparison to a DSE variant without distilla- tion ($\alpha = 0$), and to sentence embeddings that are extracted from a pre-trained BERT-Large model using the procedure described in Section 3.4.

As can be seen in Table 2, BERT-Large embeddings reach competitive results on several datasets to both In- ferSent and GenSen. Significant improvements are observed mostly for sentiment analysis related datasets. In contrast, on STS-B (semantic similarity), SICK-R, and SICK-E (NLI), BERT-Large embeddings are subpar compared to In- ferSent and GenSen, which are pre-trained directly on NLI datasets. Furthermore, recall that BERT is not explicitly trained to generate sentence embeddings, possibly explain- ing the downfalls in some of the tasks.

We now turn to compare DSE with the other baselines. As in the sentence-pair tasks evaluation, using DSE with $\alpha = 0.5$ improves upon the non-distilled variant ($\alpha = 0$), outperforming it on 8 of the 10 benchmarks. Specifically, substantial gains are obtained on SST and MRPC, demon-strating the effectiveness of knowledge distillation. There-fore, from now on, DSE relates to the $\alpha = 0.5$ model.

Dataset	Training Size	Test Size	# INV	# OOV	OOV Ratio	Source
IT-2015	11530	838	7771	1238	13.7%	Twitter trends
itter-URL	42200	9324	24905	11440	31.5%	Twitter/news
MSRP	4076	1725	16226	1614	9.0%	news

1 Model Ablations

In the original PWI model, He and Lin (2016) per-formed pattern recognition of complex semantic relationships by applying a 19-layer deep convo- lutional neural network (CNN) on the word pair interaction tensor (Eq. 5). However, the SemEval task on Interpretable Semantic Textual Similarity(Agirre et al., 2016) in part demonstrated that the semantic relationship between two sentences de- pends largely on the relations of aligned words orchunks. Since the interaction tensor in the PWI model already encodes word alignment informa- tion in the form of vector similarities, a natural question is whether a 19-layer CNN is necessary.

Table 4 shows the results of our systems with and without the 19-layer CNN for aggregating thepairwise word interactions before the final soft- max layer. While in most cases the 19-layer- CNN helps to achieve better or comparable perfor-mance, it comes at the expense of ~25% increaseof training time. An exception is the character- based PWI without language model, which per- forms well on the PIT-2015 dataset without the 19- layer CNN and comparably to logistic regressionwith string overlap features (Eyecioglu and Keller, 2015). A closer look into the datasets reveals thatPIT-2015 has a similar level of unigram overlap asthe Twitter URL corpus (Table 5),[2] but lower char- acter bigram overlap (indicative of spelling varia-tions) and lower word bigram overlap (indicativeof word reordering) between the pairs of sentencesthat are labeled as paraphrase.

Conclusion:

In conclusion, the fusion of RoBERTa and GPT in our proposed framework marks a significant advancement in the field of natural language processing. By combining the robust contextual understanding of RoBERTa with the generative capabilities of GPT, we have created a versatile model capable of excelling in a wide range of sentence-pair tasks. Through extensive experimentation and evaluation, we have demonstrated the effectiveness and scalability of our RoBERTa-GPT Fusion framework across various benchmark datasets.

Our framework not only achieves state-of-the-art performance but also offers a flexible and adaptable solution for addressing diverse NLP challenges. The synergistic integration of RoBERTa and GPT opens up new avenues for advancing attentive sentence-pair modeling, with implications for applications ranging from semantic similarity assessment to conversational AI.

Looking ahead, we envision further refinements and extensions to our framework, exploring avenues such as fine-grained task-specific adaptations, multi-modal integration, and domain-specific enhancements. By continuously pushing the boundaries of attentive sentence-pair modeling, we aim to contribute to the advancement of NLP research and pave the way for intelligent systems capable of understanding and generating natural language with unprecedented accuracy and fluency.

Computing sentence similarities via CA models such as BERT is impractical for large scale catalogs. To this end, weintroduce DSE: a sentence embedding method that is based on knowledge distillation from CA models. DSE

	Model Variations	pre-train	#parameters	Twitter URL	PIT-2015	MSRP
Word Models	Logistic Regression	–	–	0.683	0.645	0.829
	(Lan et al., 2017)	Yes	9.5M	0.749	0.667	0.834
	pretrained, fixed	Yes	2.2M	0.753	0.632	0.834
	pretrained, updated	Yes	9.5M	0.756	0.656	0.832
	randomized, fixed	–	2.2M	0.728	0.456	0.821
	randomized, updated	–	9.5M	0.735	0.625	0.834
Subword Models	C2W, unigram	–	2.6M	0.742	0.534	0.816
	C2W, bigram	–	2.7M	0.742	0.563	0.825
	C2W, trigram	–	3.1M	0.729	0.576	0.824
	CNN, unigram	–	6.5M	0.756	0.589	0.820
	CNN, bigram	–	6.5M	0.760	0.646	0.814
	CNN, trigram	–	6.7M	0.753	0.667	0.818
Subword+LM	LM, C2W, unigram	–	3.5M	0.760	0.691	0.831
	LM, C2W, bigram	–	3.6M	0.768	0.651	0.830
	LM, C2W, trigram	–	4.0M	0.765	0.659	0.831
	LM, CNN, unigram	–	7.4M	0.754	0.665	0.840
	LM, CNN, bigram	–	7.4M	0.761	0.667	0.835
	LM, CNN, trigram	–	7.6M	0.759	0.667	0.831

bypasses the need for CA operations, enabling precomputation of sen- tence representations for the existing catalog in advance, and fast query operations using a low-cost similarity func- tion. We demonstrate the effectiveness of DSE on five sen-tence-pair tasks, where it is shown to outperform other sen- tence embedding methods as well as several attentive ver- sions of ELMO. Furthermore, sentence embeddings pro- duced by DSE provide state-of-the-art results on various benchmarks.

We also showed that subword models can benefit from multi-task learn- ing with simple language modeling, and estab-lished new start-of-the-art results for paraphrase identification on two Twitter datasets, where out- of-vocabulary words and spelling variations are profound. The results shed light on future work on language-independent paraphrase identifica- tion and multilingual paraphrase acquisition where pretrained word embeddings on large corpora are not readily available in many languages.

SOME FIGURES TO ELOBARATE THE RESULTS

Dataset	Training Size	Test Size	# INV	# OOV	OOV Ratio	Source
PIT-2015	11530	838	7771	1238	13.7%	Twitter trends
Twitter-URL	42200	9324	24905	11440	31.5%	Twitter/news
MSRP	4076	1725	16226	1614	9.0%	news

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**SST**	**TREC**	**MRPC**	**SICK-R**	**SICK-E**	**STS-B**	**AVG**
GenSen	82.5	87.7	94.0	90.9	83.2	93.0	84.4/78.6	0.888	87.8	78.9/78.6	86.8
InferSent	81.1	86.3	92.4	90.2	84.6	88.2	83.1/76.2	0.884	86.3	75.8/75.5	85.2
BERT-Large	83.5	88.8	95.5	89.1	87.1	93.2	83.5/76.4	0.838	82.2	68.4/68.3	85.1
DSE ($\alpha = 0.5$)	83.6	90.2	93.6	89.8	91.0	91.8	83.8/77.9	0.856	86.7	70.7/71.4	86.4
DSE ($\alpha = 0$)	83.1	89.8	93.1	89.4	88.3	92.0	81.8/76.2	0.847	86.1	73.1/74.1	85.9

References

Barkan, O. 2017. Bayesian neural word embedding. In *AAAI*, 3135–3143.

Barkan, O., and Koenigstein, N. 2016. Item2vec: Neural item embedding for collaborative filtering. In *IEEE MLSP, 2016.*

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *EMNLP* 632–642.

Cer D.; Diab M.; Agirre E.; Lopez-Gazpio I. and Lucia S. 2017 SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval-2017*, pages 1–14, Vancouver, Canada.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12 (Aug)*:2493–2537.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 670–680.

Conneau, A., and Kiela, D. 2018. Senteval: An evaluation toolkit for universal sentence representations. *LREC*.

Dai, A. M., and Le, Q. V. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, 3061–3069.

Devlin J.; Chang M-W; Lee K; and Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Dolan, W. B. and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP@IJCNLP*.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *NIPS Workshop on Deep Learning and Representation Learning*. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.; I

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, (pp. 3294-3302).

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv*:1901.11504.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019. Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. *arXiv preprint arXiv*:1904.09482.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mnih, A., and Hinton, G. E. 2009. A scalable hierarchical distributed language model. In *NIPS*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic tex- tual similarity, English, Spanish and pilot on inter- pretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).*

Eneko Agirre, Aitor Gonzalez-Agirre, Inigo Lopez- Gazpio, Montse Maritxalar, German Rigau, and Lar- raitz Uria. 2016. SemEval-2016 task 2: Inter- pretable semantic textual similarity. In *Proceed- ings of the 10th International Workshop on Semantic Evaluation (SemEval).*

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Confer- ence on Natural Language Processing (IJCNLP).*

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large an- notated corpus for learning natural language infer- ence. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP).*

Ohio Supercomputer Center. 2012. Oakley supercomputer. http://osc.edu/ark:/19495/hpc0cvqn.

Marta R. Costa-jussa` and Jose´ A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Associa-tion for Computational Linguistics (ACL).*

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING).*

William B Dolan and Chris Brockett. 2005. Automati-cally constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP).*

Cicero dos Santos and Victor Guimarães. 2015. Boost- ing named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named En- tity Workshop*.