Attentive Siamese LSTM Network for Semantic Textual Similarity Measure

Wei Bao*‡, Wugedele Bao†, Jinhua Du‡, Yuanyuan Yang§ and Xiaobing Zhao*
*National Language Resource Monitering & Research Center of Minority Languages
Minzu University of China, Beijing, China 100081
Email: jsnubw@163.com, nmzxb_cn@163.com
†Hohhot Minzu University, Hohhot, Inner Mongolia Autonomous Region, China
Email: wgd2827@163.com
‡Dublin City University,Dublin, Ireland
Email: jinhua.du@adaptcentre.ie
§Bohai University, Jinzhou, Liaoning
Email: wateryoo919@aliyun.com
Corresponding author: Xiaobing Zhao

***Abstract*—Semantic Textual Similarity (STS) is important for many applications such as Plagiarism Detection (PD), Text Paraphrasing and Information Retrieval (IR). Current methods for STS rely on statistical machine learning. Recent studies showed that neural networks for STS presented promising experimental results. In this paper, we propose an Attentive Siamese Long Short-Term Memory (LSTM) network for measuring Semantic Textual Similarity. Instead of external resources and handcraft features, raw sentence pairs and pre-trained word embedding are needed as input. Attention mechanism is utilized in LSTM network to capture high-level semantic information. We demonstrated the effectiveness of our model by applying the architecture in different tasks: three corpora and three language tasks. Experimental results on all tasks and languages show that our method with attention mechanism outperforms the baseline model with a higher correlation with human annotation.**

Keywords-semantic textual similarity; attention mechanism; siamese LSTM;

I. Introduction

STS seeks to "measure the degree of semantic equivalence between two sentences/fragments", which can be used in many NLP applications, such as Plagiarism Detection (PD), Information Retrieval (IR) and many other tasks.

Plagiarism is a serious academic misconduct, which refers to the "wrong appropriation" and "stealing and publication" of the original texts as one's own original work without appropriate citations. Zhang analyzed 662 scientific papers using a PD tool: 22.8% of these papers contained apparently unreasonable levels of copy-paste plagiarism in which 25.8% of these papers were serious plagiarism. Similarity of some cases was as high as 83% [1]. ACM and IEEE has built the guidelines to avoid scientific misconduct. There are many plagiarism detection tools available online now as described in [2]. Nearly all tools focus on string matches and fingerprint to detect copy-paste and verbatim plagiarism, which is easy to be detected. While little tools can detect paraphrasing and semantic plagiarism, which is a quite challenging task due to the fact that semantic information in paraphrased sentence is hard to be extracted.

To deal with this issue, Gipp showed that even strongly paraphrased texts remained similar order of citations in a paper, so they proposed a Citation-based Plagiarism Detection (CbPD) method to detect paraphrasing and semantic plagiarism[3][4]. CbPD implemented four algorithms to analyze citation pattern, including "greed citation tiling", "longest common citation sequence", "citation chunking" and "adaptive bibliographic coupling". The main disadvantage of citation-based method is its dependence on correct citations, so papers with incorrect citations or few citations are hard to be detected. In this paper, we aim at content-based plagiarism detection, in which the first task is semantic textual similarity measure.

Deep neural networks (DNNs) have made significant progress and become state-of-the-art (SOTA) in many NLP tasks, such as question answering, machine translation etc. [5][6][7]. Due to the end-to-end method is SOTA, we propose an attentive Siamese LSTM architecture to measure semantic sentence pairs. Three tasks are used to demonstrate the performance of our method, including SemEval semantic relatedness task [8], Microsoft Research paraphrase identification task [9] and Chinese Mandarin and Tibetan corpora translated from SemEval task. Experimental results show that Attentive Siamese LSTM architecture is more effective than the baseline model in all tasks above.

Structure of this paper is as follows: we make an elaborate review of related work of STS in Section II. Section III presents the methodology used in our system. Experimental results and analysis are discussed in section IV. Finally, we have conclusions and future work in section V.

II. Related Work

SemEval is an important evaluation for semantic similarity held by ACL. English sentences semantic relatedness was the primary task in SemEval2014 and has achieved great success. The top-ranking system in SemEval2014 used compositional features or non-compositional features with external resources, such as WordNet. Numerous heterogeneous features with priori knowledge were

978-1-7281-1175-9/18/$31.00 ©2018 IEEE

utilized, e.g. word similarity/overlap, syntactic features, sentence/phrase composition and negation features. Many learning methods were used in the system, e.g. SVM, KNN, Random Forest and combination classifiers [8]. Zhao et al. used 7 types of features with a regression model to make prediction of sentence similarity [10]. To obtain better performance, 72 features were extracted from the sentence pair, including the surface text similarity, semantic similarity, grammatical relationship and corpus-based feature and so on. Another top-ranking work presented a method using high-level language knowledge, used formal semantics and logical inference along with 32 handcrafted-features to predict semantic similarity [11].

Recently, deep neural networks have representative progress in semantic textual similarity. He et al. proposed multi-prospective convolutional neural network (MPCNN) [12] and Tai et al. presented tree-structured LSTM network topology to model sentence pairs [13]. Different from the standard LSTM model, tree LSTM model can encode semantically-useful structural information in the sentence. Skip-thought vectors model was proposed by Kiros et al. [14] to further improve model performance.

Neural networks have achieved huge success in STS task as well, so we choose the basic end-to-end model as our baseline system. Considering various lengths of sentences in given pairs, we propose an Attentive Siamese LSTM architecture to capture important semantic information in sentence pair.

The contributions of this paper are as follows: (1) we propose a novel architecture for STS task, i.e. an Attentive Siamese LSTM, which can learn underlying semantic information in a sentence; (2) we conduct comparable experiments on different tasks, i.e. SemEval2014 task 1 corpus, additional Chinese Mandarin and Tibetan corpus and MSRP corpus. Strong and moderate correlation between predicting similarity and human annotated similarity in two SemEval corpora are achieved. Results of all tasks indicate that our Attentive Siamese LSTM model outperforms the baseline model.

III. Methodology

In this section, we describe Siamese LSTM architecture firstly and then propose Attentive Siamese LSTM model in Figure 1 in detail. Five layers in this model are described. Finally, word embeddings used in our experiments are introduced.

A. Siamese LSTM Network

Siamese architecture has been utilized in many NLP applications, such as vision application[15][16][17] and acoustic modelling[18][19], and has achieved great success in image, speech similarity measure. More recently, Siamese architecture has been applied to measure text similarity[20].

(Mueller and Thyagarajan, 2016) presented Siamese LSTM architecture to learn sentence semantic similarity, which obtained better performance than the other methods. Siamese architecture has two identical sub-networks and each processes a sentence in the given pair, which is especially suitable for similarity measure.

LSTM (Hochreiter and Schmidhuber, 1997) is designed to cope with the backpropagated gradients vanishing problem of basic RNN. A LSTM unit consists three multiplicative gates, an input gate, a memory gate and an out gate, these three gates will help LSTM learn long range dependencies. Formulas updating a LSTM unit at time t are as follows:

$$i_t = \sigma(W_i h_{i-1} + U_i x_t + b_i) \tag{1}$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{2}$$

$$\widetilde{c} = tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{3}$$

$$c_t = f_i c_{t-1} + i_t \widetilde{c_t} \tag{4}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{5}$$

$$h_t = o_t tanh(c_t) \tag{6}$$

x_t is the input vector and h_t is the hidden state at time t. W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o is the weight matrices and b_i, b_f, b_c, b_o is the bias vector.

Siamese LSTM architecture in (Mueller and Thyagarajan, 2016) has three layers. There are two LSTM networks where each process one of the sentences of the input sentence pair. Two LSTM share the same weights in Siamese architecture. Bidirectional LSTM model[21][22] utilize a second layer which has the opposite order with the first layer to combine future context and past context. We also adopt this model in this paper to conduct comparable experiments.

Training data $(x)_1^{(a)}, \cdots, (x)_{T^a}^{(a)}, (x)_1^{(b)} \cdots, (x)_{T^b}^{(b)}$, is of fixed-size vector with a label y presenting for semantic similarity. a and b indicates one sentence in the given pair. T_a, T_b indicates sentence length, which may be different length. Similarity function $g = exp(-\left\|(h)_{T_a}^a - (h)_{T_b}^b\right\|_1) \in [0,1]$ is computed by the final hidden state of the model. To obtain predicting similarity$\in [1,5]$, we rescale g lie $\in [1,5]$.

In (Mueller and Thyagarajan, 2016),only the last hidden states $h_4^{(a)}$, $h_5^{(b)}$ are utilized, while sentence in one pair may be of different length. So we propose Attentive Siamese LSTM network to capture underlying information hidden in the sentence pair.

B. Attention Mechanism

Neural networks with attention mechanism have achieved great success in many NLP tasks[23][24][25][26]. In this paper, we proposed Attentive Siamese LSTM architecture for semantic textual similarity measure.

Siamese LSTM model use the last hidden state of sentence pair $h_4^{(a)}$, $h_5^{(b)}$, which will ignore many information. We use all hidden state $H_a = [h_1^{(a)}, h_2^{(a)}, \cdots h_3^{(a)}, h_{T_a}^{(a)}]$, $H_b = [h_1^{(b)}, h_2^{(b)}, \cdots h_3^{(b)}, h_{T_b}^{(b)}]$ as input, where T_a, T_b is the length of sentence pair. α is the weight of $LSTM_a$ and β is the weight of $LSTM_b$. The final representation of sentence pair is r_a, r_b computed by equation 9, which

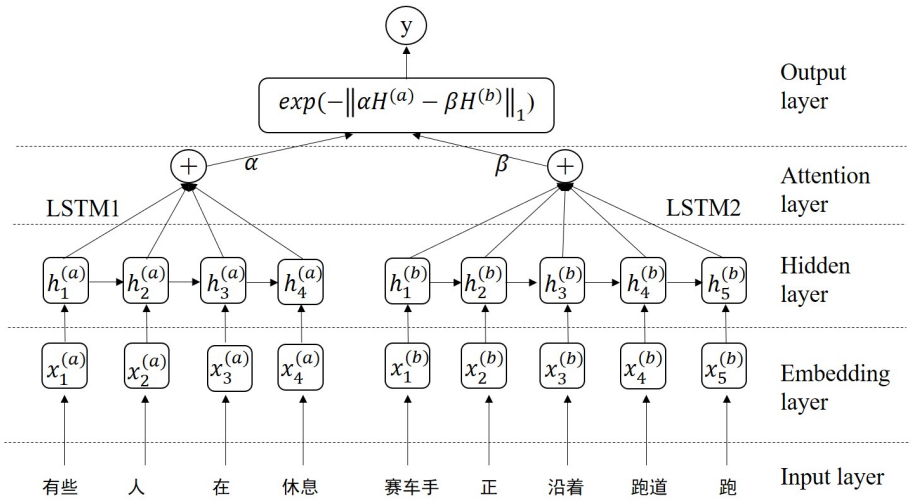

Figure 1. General Structure of Our Siamese LSTM Network with Attention Mechanism (AttSiaLSTM).

is a weighted sum of the output vectors from equation 7, equation 8.

$$M = tanh(H) \tag{7}$$

$$\alpha = softmax(wM) \tag{8}$$

$$r = H\alpha \tag{9}$$

Attentive Siamese LSTM model proposed in our paper is shown in Figure 1, which contains two networks, $LSTM_a$ and $LSTM_b$, which share the same weight. $LSTM_a$ and $LSTM_b$ process one of the sentence in a given pair. Each network has five layers, 1) input layer: input the given sentence pair, 2) embedding layer: represent words in a low dimension, 3) hidden layer: learn high-level features, 4) attention layer: produce weight vector, 5) output layer: output predicting similarity (or label). Instead of inputting numerous of handcraft features, sentence pair with similarity and word embedding is needed in our model.

C. Word Embedding

Instead of numerous manually features, word embedding is the only feature in our system. In this paper, we use English and Chinese 300-dimensional word embeddings, which is pretrained and publicly available. These monolingual word vectors are trained on Wikipedia corpus using fastText in 300-dimension[27] . English word embedding and Chinese word embedding is 6.1GB and 821MB. Tibetan word embedding available online is only 32.1MB, containing 12651 words. So we train a big-scale Tibetan word embedding using corpus collected from Tibetan news websites and Tibetan scientific publications and so on.

IV. Experiments and Analysis

In this section, we conduct a series of experiments to test the model's performance in different corpus and different language. Experiments of different sentence length and bilingual STS task are explored. Description of STS experimental set wiil be given. Experimental results and analysis will be shown at last.

sen1_En	A couple of policewomen are singing.
sen2_En	Two women are dancing.
sen1_Zh	一对女警正在唱歌。
sen2_Zh	两个女人在跳舞。
sen1_Ti	བྱད་མེད་ཞེན་རྒོག་པ་ཚ་གཅིག་གླུ་ལེན་བཞིན་ཡོད།
sen2_Ti	བྱད་མེད་གཉིས་ཀྱིས་ཨང་སོ་འཁྱབ་བཞིན།
sen1_En	A dirty soccer ball is rolling into a goal net.
sen2_En	A soccer ball is rolling into a goal net.
sen1_Zh	一个肮脏的足球正在滚进一个球门网。
sen2_Zh	一个足球滚进一个球门网。
sen1_Ti	བཙོག་པོ་ཞིག་གི་རྐང་རྡེད་རྡོ་ལོའི་བཞིན་ཡོད་པ་ཞིག་རྒྱིས།
sen2_Ti	རྐང་རྡེད་རྡོ་ལོ་ཞིག་ཞིག་རྒྱིས།

Figure 2. Sample of Corpus in English, Chinese Mandarin and Tibetan.

A. Corpus

There are little corpora available for STS task, low-resource corpora for STS is scarce. In this paper, we consider three sentence similarity tasks.

1) SemEval2014 evaluation corpus. Semantic Evaluation (SemEval) is an important evaluation for STS task.The Sentences Involving Com-positional Knowledge (SICK) data set[8] released in SemEval 2014 is to evaluate semantic relatedness between the sentence pair. It consists of 10,000 English sentence pairs, [training, 4500], [development, 500], [test, 5000]. Each sentence has an similarity label $\in$[1, 5], with 1 indicating that two sentences in a given pair are completely unrelated while 5 indicating that two sentences are completely related.
2) SemEval2014 translated corpus. In addition, English annotated sentence pairs were translated into Chi-

Table I
EXPERIMENTAL RESULTS OF DIFFERENT CORPUS IN ENGLISH, CHINESE MANDARIN AND TIBETAN. CORPUS.

Model	EN			ZH			TIB			MSRP	
	r	Increase	MSE	*r*	Increase	MSE	*r*	Increase	MSE	*p*	Increase
Baseline1	0.7121		0.032	0.328		0.0598	0.3451		0.0519	60.41%	
Baseline2	0.6326		0.0388	0.3062		0.0570	0.2854		0.0867	61.68%	
AttSiaLSTM	0.7832	↑ 0.0711	0.026	0.4646	↑ 0.1584	0.0428	0.3627	↑ 0.0773	0.0813	65.68%	↑ 5.27%
AttSiaBiLSTM	0.7518	↑ 0.1192	0.031	0.4173	↑ 0.0553	0.0386	0.3437	↑ 0.0583	0.0782	63.19%	↑ 1.51%

Table II
EXPERIMENTAL RESULTS OF DIFFERENT SENTENCE LENGTH

Language	Method	Length	*r*	MSE
Zh	SiaLSTM	15	0.2919	0.0592
Zh	SiaLSTM	20	0.3062	0.0570
Zh	AttSiaLSTM	15	0.4334	0.0452
Zh	AttSiaLSTM	20	0.4646	0.0428
Zh	AttSiaLSTM	25	0.4547	0.0459
Tib	SiaLSTM	15	0.2325	0.0903
Tib	SiaLSTM	20	0.2585	0.0878
Tib	SiaLSTM	25	0.3451	0.0519
Tib	AttSiaLSTM	15	0.3258	0.0836
Tib	AttSiaLSTM	20	0.3566	0.0816
Tib	AttSiaLSTM	25	0.3627	0.0812

nese and Tibetan to create additional tracks in our experiments, which will be used for Chinese Mandarin and Tibetan Plagiarism Detection. Chinese sentences are translated from English via Google Translator[1] and Tibetan sentences are translated from Chinese via Niu Translator[2]. For each language we got 10,000 sentence pairs in total. Two samples in English, Chinese Mandarin and Tibetan are shown in Figure 2. Related score of two samples are 1.7 and 4.5.

3) MSRP. Another paraphrasing corpus is Microsoft Research Paraphrase corpus (MSRP), including 5801 English sentence pairs with human annotated label. Training set contains 4076 sentence pairs and test set contains 1725 sentence pairs. 67% sentence pairs are paraphrasing sample and 33% are not.

B. Experimental Results

Our Siamese LSTM network uses 50-dimensional hidden representations. Word embeddings used in our experiments is 300-dimension. We use the official evaluation metric to evaluate our method. Pearson correlation coefficients r and Mean-square error MSE are used to evaluate the performance of SemEval corpus, which indicates the difference between the sentence pair's predicting similarity and human annotated similarity. Precision is used to evaluate the performance of MSRP corpus. Equations are as following.

$$r = \frac{Cov(sim_{out}, sim_{label})}{\sqrt{Var(sim_{out})Var(sim_{label})}} \tag{10}$$

$$MSE = sum((sim_{out} - sim_{label})^2) \tag{11}$$

[1]https://translate.google.cn
[2]http://fanyi.niutrans.com/

Pearson correlation coefficient ranges from -1 to 1, which implies negative correlation and positive correlation. Value of [0.8,1.0], [0.6,0.8], [0.4,0.6], [0.2,0.4], [0,0.2] represents "perfect correlation", "strong correlation", "moderated correlation", "weak correlation" and "zero correlation". Experimental results are in Table I.

Below we compare the performance of attentive Siamese LSTM network, against the baseline model without attention mechanism. Table 1 shows the results of experiments. 'Baseline1' is the results of Siamese LSTM model while 'Baseline2' is the results of Siamese bidirectional LSTM model. '*r*' implies Pearson correlation and 'Increase' is the increase of attentive model than the corresponding baseline model.

The first and most important thing to note is that in both SemEval and MSRP corpus, all of the attentive models showed better performance than the corresponding baseline models without any handcraft features and external sources. Pearson's correlation of English sentence pair can reach 0.7832, which indicates predicting similarity and human annotated similarity is strong correlated, near perfect correlated. In MSRP task, AttSiaLSTM and AttSiaBiLSTM model gain5.27% and 1.51% improvement than the corresponding baseline model. It proves that attentive Siamese LSTM model can capture important semantic information in a sentence, which is effective for our STS task.

We note that Siamese LSTM model obtain better performance than Siamese bidirectional LSTM model. This conclusion is in keeping with (Mueller and Thyagarajan, 2016). So we choose Siamese LSTM model to conduct our experiments.

We can see from Table I, Pearson correlation of Chinese Mandarin and Tibetan show the same tendency with English: attentive model achieves better results. Attentive Siamese LSTM of Chinese Mandarin shows 0.1584 improvement over the baseline model while Tibetan experiment shows 0.0773 improvement. This indicates that the proposed method with attention mechanism is suitable for learning sentence similarity.

In addition, among three languages in Table I, English achieves the best results and a large drop in Pearson correlation on the Chinese Mandarin and Tibetan sentence pair. We think one reason of this performance gap is the scale of word embedding. As showed above, English word embedding is 6.1GB while Chinese Mandarin and Tibetan word embedding is 821MB, 255MB, words in which are 2,519,270, 332,647, 102,054. Scale and domain of word embeddings do affect the performance of the model, which

can been seen from two types of Tibetan embeddings experiments elaborated later. For lack of Chinese Mandarin and Tibetan STS corpus, we used translation corpus in this experiments, which may bring errors by translation quality. In future, we will collect real Chinese Mandarin and Tibetan STS corpus for STS task.

Two Tibetan word embeddings were used in the experiments, small-scale one is available online from Facebook, another is trained by ourselves (as showed in Table II). Pearson correlation of the small-scale one is 0.3456, which has 0.1102 improvement against the model without attention mechanism. Compare with 'TIB' results in Table I, we can see that, large-scale Tibetan word embeddings have better performance. This gain is to be expected since we used a large-scale Tibetan word embeddings.

C. Analysis and Discussion

To test whether sentence length do affect the performance, we proposed a series of experiments as shown in Table II. We treated max sentence length as a parameter to be tuned and experimental results showed it could improve performance. Chinese Mandarin task achieved a correlation of 0.4646 with the length 20 model while Tibetan task achieved a correlation of 0.3627 with the length 25 model, which really do affect the performance of the model. This suggests that we can tune sentence length as a parameter to gain better results.

In addition, we use English-Chinese translated, parallel sentence pairs to create additional bilingual STS task in our experiments, which is also an important task for cross-lingual plagiarism detection. We use two bilingual word embedding provided by Facebook. MUSE[28][29] provides two methods to train a bilingual word embedding: supervised method with bilingual dictionary and unsupervised machine translation with monolingual data only. We use supervised method to train an English-Chinese bilingual word embedding. English-Chinese bilingual dictionary is publicly available online provided by Facebook[3]. The bilingual embedding is trained using the default parameters and with a dimension of 300.

Another English-Chinese bilingual word embedding published by Facebook[4] is applied in our experiments, which aligns monolingual vectors from two languages in a single vector space by using SVD to learn a linear matrix[30]. Results show that predicting similarity of our model is moderate correlated with human annotated similarity, which is a benefit work for cross-lingual semantic textual similarity measure and plagiarism detection.

V. Conclusion

In this paper, Attentive Siamese LSTM network is proposed to measure semantic textual similarity. Instead of numerous manually features rely on priori knowledge or external resources, our model utlize raw sentence pair and pre-trained word embeddings as input. Experimental results in three tasks and three languages show that our model with an attention mechanism is effective in STS tasks, which will be beneficial to Plagiarism Detection. We also explore English-Chinese bilingual STS task, which obtain a moderate related result between predicting similarity and human annotated label.

In future we will focus on two works, the first one is to collect real Chinese Mandarin and Tibetan STS corpus to test robustness of our Attentive Siamese LSTM method. Second work is to explore bilingual and multilingual STS tasks to lay foundation for bilingual and multilingual Plagiarism detection.

[3]https://github.com/facebookresearch/MUSE

[4]https://github.com/Babylonpartners/fastTextmultilingual

Acknowledgment

This paper was supported by National Natural Science Foundation of China (NSFC) (No.61331013,No.61501529), National Language Commission (ZDI135-39), National Social Science Foundation (No.17CYY044).

References

[1] H. Zhang, "Crosscheck: An effective tool for detecting plagiarism," *Learned Publishing*, vol. 23, pp. 9–14, 2010.

[2] L. S. Pertile, P. V. Moreira, and P. Rosso, "Comparing and combining content- and citation-based approaches for plagiarism detection," *Journal of the Association for Information Science and Technology*, vol. 67, no. 10, pp. 2511–2526, 2016.

[3] G. Bela and M. Norman, "Citation pattern matching algorithms for citation-based plagiarism detection: Greedy citation tiling, citation chunking and longest common citation sequence," in *Proceedings of the 11th ACM Symposium on Document Engineering*, ser. DocEng '11, 2011, pp. 249–258.

[4] G. Bela, M. Norman, and B. Corinna, "Citation-based plagiarism detection: Practicability on a large-scale scientific corpus," *Journal of the Association for Information Science and Technology*, vol. 65, no. 8, pp. 1527–1540, 2014.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model." in *Procceddings of INTERSPEECH*. ISCA, 2010, pp. 1045–1048.

[6] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation." in *Procceddings of Empirical Methods in Natural Language Processing*. ACL, 2013, pp. 1393–1398. [Online]. Available: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2013.htmlZouSCM13

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch." *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/jmlr/jmlr12.htmlCollobertWBKKK11

[8] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation@COLING*, 2014.

[9] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. [Online]. Available: http://www.aclweb.org/anthology/I05-5002

[10] J. Zhao, T. Zhu, and M. Lan, "Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation@COLING*, 2014.

[11] J. Bjerva, J. Bos, R. van der Goot, and M. Nissim, "The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity," in *Proceedings of the 8th International Workshop on Semantic Evaluation@COLING*, 2014.

[12] H. He, K. Gimpel, and J. J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Procceddings of Empirical Methods in Natural Language Processing*, 2015, pp. 1576–1586.

[13] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. abs/1503.00075, 2015, p. 1556–1566. [Online]. Available: http://arxiv.org/abs/1503.00075

[14] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," *CoRR*, vol. abs/1506.06726, 2015. [Online]. Available: http://arxiv.org/abs/1506.06726

[15] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06, 2006, pp. 1735–1742.

[17] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," *CoRR*, vol. abs/1412.6622, 2014. [Online]. Available: http://arxiv.org/abs/1412.6622

[18] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proceedings of INTERSPEECH*, 2015, pp. 3179–3183.

[19] G. Synnaeve and E. Dupoux, "A temporal coherence loss function for learning unsupervised acoustic embeddings," *Procedia Computer Science*, vol. 81, pp. 95 – 100, 2016.

[20] P. Neculoiu, M. Versteegh, and M. Rotaru, "Similarity with siamese recurrent networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP*, 2016, p. 148–157.

[21] D. Zhang and D. Wang, "Relation classification via recurrent neural network," *CoRR*, vol. abs/1508.01006, 2015. [Online]. Available: http://arxiv.org/abs/1508.01006

[22] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *CoRR*, vol. abs/1508.01991, 2015. [Online]. Available: http://arxiv.org/abs/1508.01991

[23] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. MIT Press, 2015, pp. 1693–1701. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969428

[24] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2016, pp. 207–212.

[25] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 577–585. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969304

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015. [Online]. Available: http://arxiv.org/abs/1502.03044

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016. [Online]. Available: http://arxiv.org/abs/1607.04606

[28] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *CoRR*, vol. abs/1710.04087, 2017. [Online]. Available: http://arxiv.org/abs/1710.04087

[29] G. Lample, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *CoRR*, vol. abs/1711.00043, 2017. [Online]. Available: http://arxiv.org/abs/1711.00043

[30] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *CoRR*, vol. abs/1702.03859, 2017. [Online]. Available: http://arxiv.org/abs/1702.03859