PHRASE-LEVEL GLOBAL-LOCAL HYBRID MODEL FOR SENTENCE EMBEDDING

Mingyu Tang[*†], *Liansheng Zhuang*[*†], *Houqiang Li*[*], *Jian Yang*[†‡], *Yanqun Guo*[§]

[*]University of Science and Technology of China, [†]Peng Cheng Laboratory
[‡]Northern Institute of Electronic Equipment, [§]Southwest Jiaotong University
tmy528@mail.ustc.edu.cn, lszhuang@ustc.edu.cn

ABSTRACT

Latent structure models have drawn much attention due to the ability to learn an optimal latent hierarchical structures without explicit structure annotations. However, most existing models suffer from high computation complexity and hard training. To this end, this paper proposes a novel phrase-level global-local hybrid model, which inherits the advantages of existing latent structure models while requires less time complexity. Our model splits a sentence into multiple phrases by a category-selection module. Then, it encodes the context dependency by a phrase-level global encoding module, and encodes the task-specific information by a phrase-level local encoding module. Finally, sentence embedding is obtained by integrating the global encoding and task-specific encoding. Experiments on public benchmarks show that, our model achieves state-of-the-art performance on the tasks of sentence classification and natural language inference. Meanwhile, our model is at least 10 times faster than existing state-of-the-art method at the training stage.

Index Terms— sentence embedding, latent structure model, phrase level

1. INTRODUCTION

Sentence embedding plays a critical role in many natural language processing (NLP) applications such as sentiment analysis [1], question answering [2] and entailment recognition [3]. It aims at mapping sentences into dense real-valued vectors that represent their semantics. Though word embedding has achieved great success [4], there is no clear way to build embedding that carries full meaning of a sentence, which is the basic processed unit in most downstream tasks.

Previous methods for sentence embedding mainly rely on Recurrent Neural Networks(RNNs) to generate context-aware representation [5]. Though RNNs encoders are capable of learning long-term dependencies by reading words in sequential order, they are hard to parallelize and not time-efficient. Beyond that, attention-based methods [6] use attention mechanism to build representations by scoring input words differentially. Beyond that, attention-based methods [6] use attention mechanism to build representations by scoring input words differentially, which can be more parallelizable and requires significantly less time to train than RNNs. Both RNNs models and attention-based models have achieved striking performance, but they can not well encode linguistic information of natural language which can impact their perform to some extent. Language is inherently tree structured, and the meaning of a sentence comes largely from composing the meanings of sub-trees. To explicitly model the composition, tree-based models [7, 8] are proposed to use pre-specified parsing trees to embed sentences by recursive manner. Although there is significant benefit in processing a sentence in a tree-structured, data annotated with parse trees could be expensive to prepare and hard to be computed in batches which seriously affect training efficiency.

Recently, latent tree models have been proposed to learn the optimal hierarchical latent structure of text from sequence into sentence representation without specific tree annotation [9]. The training signals to parse and embed sentences are both from certain downstream tasks. However, most existing latent structure models suffer from time consuming problems and hard training due to the ineffective reinforcement learning method [10] or heavy calculation for candidate parent tree nodes [11]. Meanwhile, most of these models still follow the way of parsing sentence to binary tree with words in leaf nodes and then composing adjacent node pairs bottom up. This prevents the sentence embedding from focusing on the most informative words, resulting in a performance limitation on certain tasks [12]. We argue that weak form of syntax tree can be better for sentence embedding by neural network.

Motivated by above insights, this paper proposes a novel phrase-level global-local hybrid(PLGLH) model, which inherits the advantages of existing latent structure models with less time complexity. Our model splits a sentence into multiple phrases by a category-selection module rather than reinforce learning method like HS-LSTM [11], and obtains a weak form of constituent in the syntax of a sentence. To filter out information that is semantically or syntactically distant, each phrase is encoded respectively. To capture the context

This work was supported in part to Dr. Liansheng Zhuang by NSFC under contract No.61976199, and in part to Dr. Houqiang Li by NSFC under contract No.61836011. Dr. Liansheng Zhuang is the corresponding author.

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.

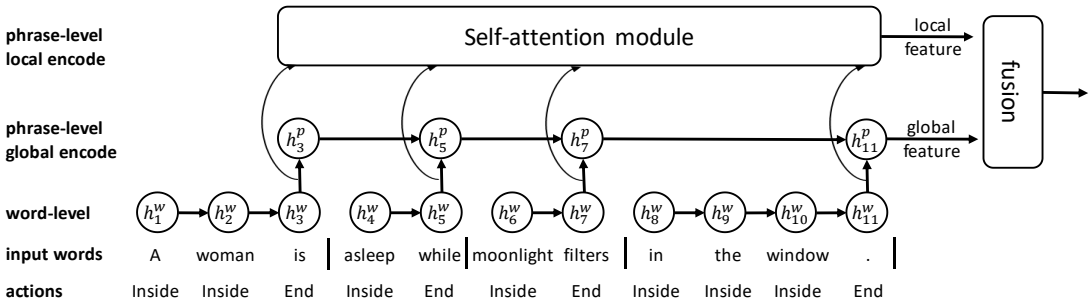

Fig. 1. Overall process of proposed model

dependencies among phrases, a phrase-level global encoding module is introduced based on LSTM, and process these phrases by sequence. Since different phrases have different relevance to the task, our model introduce a phrase-level local encoding module based on attention mechanism to emphasize the task-informative phrases. The final sentence embedding is obtained by fusing the global encoding and local encoding, and then applied in a downstream application. Compared with existing latent tree models which are based on reinforce learning or recursive composing, our model is easy to train, and significantly reduces the training time.

We evaluate our model's performance on a plethora of datasets of natural language inference and sentence classification. Experiment results show that our model outperforms or is at least comparable to previous sentence encoder models.

The main contributions of our work can be summarized as follows:

- we propose a novel phrase-level global-local hybrid model for sentence embedding that can combines global context dependency and task-specific local information.
- Our model inherits the advantages of existing latent structure with less time complexity.
- Our proposed model outperforms the state-of-the-art supervised sentence embedding method on a wide range of datasets.

2. PROPOSED MODEL

In this section, we introduce the PLGLH model for sentence encoding. Formally, given a sentence $X = (x_1, x_2, ..., x_T)$ with length of T, where x_t represents the word embedding of corresponding word, our aims is to build a fix size vector emb to summarize the information of it. The overall process is shown in Figure 1.

2.1. Overview

First, our model applies a word-level module to split a sentence into multiple phrases by Straight-Trough (ST) Gumbel-Softmax estimator [13] rather than similar HS-LSTM which uses reinforce learning method, and obtains a weak form of constituent in the syntax of a sentence. To filter out information that is semantically or syntactically distant, each phrase is encoded respectively, detail in section 2.2. To capture the context dependencies among phrases, a phrase-level global encoding module is introduced based on LSTM, and process these phrases by sequence, detail in section 2.3. Since different phrases have different relevance to the task, our model introduce a phrase-level local encoding module based on attention mechanism to emphasize the task-informative phrases, detail in section 2.4. The final sentence embedding is obtained by fusing the global encoding and local encoding.

2.2. Word-level Phrase Division

Our model splits sentence into phrases and build representation for them in this phase. In particular, it is formulated as a sequence decision by action $a_t \in \{Inside, End\}$. When a_t is $Inside$ means x_t is the start of one phrase, or belongs to same phrase as x_{t-1}, otherwise a_t is End means x_t is the end of that phrase. As result, words in sentence will be split into a few of consecutive parts, or called phrase.

We use a word-level LSTM to connect a sequence of words to form a phrase representation. The transition of the word-level LSTM depends upon action a_{t-1}. If action a_{t-1} is End, the word at position t is the start of a phrase and the word-level LSTM start with a initialized state, which also can be trained in the training phase. Otherwise the action is $Inside$ and the world-level LSTM continues from its previous states. The process is described formally as follows:

$$c_t^w, h_t^w = \begin{cases} \Phi^w(x_t, c_0^w, h_0^w), & a_{t-1} = End \\ \Phi^w(x_t, c_{t-1}^w, h_{t-1}^w), & a_{t-1} = Inside \end{cases} \quad (1)$$

where Φ^w denotes the transition function of word-level

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.

LSTM, c_t is the memory cell, and h_t is the hidden state at timestep t.

Since the action of phrase division is not given to the model, we need to build up the structure at the same time. We choose the ST Gumbel-Softmax estimator [13] to achieve sampling from discrete action space. It is a method of utilizing discrete random variables in a network by approximating one-hot vectors sampled from a categorical distribution. Consider the k-dimensional categorical distribution whose class probabilities $p_1, p_2, ..., p_k$ are defined by unnormalized log probabilities$\pi_1, \pi_2, ..., \pi_k$, as show in equation 2. Then ST Gumbel-Softmax is processed as equations 3.

$$p_i = \frac{exp(log(\pi_i))}{\sum_{j=1}^{k} exp(log(\pi_j))} \tag{2}$$

$$\begin{aligned} u_i &\sim Uniform(0,1) \\ g_i &= -log(-log(u_i)) \\ y_i &= \frac{exp((log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} exp((log(\pi_j) + g_j)/\tau)} \\ y_i^{ST} &= \begin{cases} 1, & i = arg\,max_j y_j \\ 0, & otherwise \end{cases} \end{aligned} \tag{3}$$

where u_i is random noise; τ is temperature parameter and be set to 1 in our experiment; $y = (y_1, y_2, .., y_k)$ is similar as standard softmax output and can be used for backpropagation; $y^{ST} = (y_1^{ST}, y_2^{ST}, ..., y_k^{ST})$ is one-hot vector that can represent discrete sampling in forward propagation.

Come back to our model, we use linear transformation to map the word embedding and hidden state to two-dimensional action space, represented as (π_1, π_2), and then use ST Gumbel-Softmax to sample the action from it at every time step.

The output a_t at forward propagation will be $(1, 0)$ or $(0, 1)$, representing the action of End or $Inside$.

Finally we will get a sequence of outputs of hidden states and actions, $(h_1, h_2, ..., h_T)$ and $(a_1, a_2, ..., a_T)$, where T is the length of input words. When $a_t = (1, 0)$, it indicates that h_t represents the meaning of some phrase. Otherwise, h_t is invalid and won't be used in later. But in order to support batched computation, we won't throw away the invalid part of h directly.

After that, each phrase has been encoded, and we can use them to build the embedding for sentence.

2.3. Phrase-level Global Encoding

Follow the common practice in latent structure model, we use another phrase-level LSTM to capture the context dependency of whole sentence from sequence of phrases.

As we can see in Figure 1, actions produced in word-level module will guide us to select valid hidden state. When the action a_t is End, a phrase ends at position t and the hidden state h_t which carrying means of it will be fed into phrase-level LSTM. Otherwise the action is $Inside$ and the phrase-level LSTM is fixed at this step, and the variables are copied from the preceding position. Formally as follow:

$$c_t^p, h_t^p = \begin{cases} \Phi^p(h_t^w, c_{t-1}^p, h_{t-1}^p), & a_{t-1} = End \\ c_{t-1}^p, h_{t-1}^p, & a_{t-1} = Inside \end{cases} \tag{4}$$

where Φ^p denotes the transitions function of phrase-level LSTM, h_t^w is the output of word-level LSTM. The last output hidden state of phrase-level LSTM will be considered as the global encoding for the sentence.

2.4. Phrase-level Local Encoding

Since different phrases have different relevance to specific task, our model introduce a phrase-level local encoding module based on attention mechanism to emphasize the task-informative phrases. We follow the Scaled Dot-Product Attention of Transformer [14].

First, filter the invalid hidden state. Different with global encoding with LSTM, we can just force the invalid item to zero. Then map each valid hidden state to query(Q), key(Key) and value(V) simultaneously. Inner product of key and query can indicate the relevance between two phrases. New summarize representation for each time step is the weighted sum of each part's value which simulates the relevance as each part can exchange information directly and help to find the important message. The whole process can be seen in Fig 2, formally in matrix form as follows:

$$\begin{aligned} \hat{h}^w &= ActionMask(h^w, a) \\ Q, K, V &= Linear(\hat{h}^w) \\ Z &= softmax(\frac{QK^T}{\sqrt{d_k}})V \\ Z_{out} &= MeanPooling(Z) \end{aligned} \tag{5}$$

where $ActionMask$, as the name, use the actions generated in word-level module to mask the h^w to force the invalid parts to zero. $Linear$ denotes the linear transformation without bias, which keeps the zero items still be zero and transfers hidden state to query, key and value vectors separately. d_k, which denote the dimension of Q and K, is used as scaling factor. Mean-pooling is used to map the attention module output to fix size embedding.

Finally, we can fusion two embeddings to get the sentence representation compressed with local and global information. In this paper, we just simply concatenate two embeddings as the final result to keep information as much as possible.

$$emb = h_T^p \oplus Z_{out} \tag{6}$$

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.

Table 1. Dataset examples

name	N	example	label
SNLI	550k	Two women are embracing while holding to go packages.(premise) Tow woman are holding packages. (hypothesis)	entailment
MultiNLI	433k	I don't know um do you do a lot of camping.(premise) I know exactly. (hypothesis)	contradiction
MR	11k	Unfortunately the story and actors are served with a hack script.	neg
CR	4k	This is by far the nicest one, in so many ways.	pos
SUBJ	10k	Smart and alert, thirteen conversations about one thing is a small gem.	sub

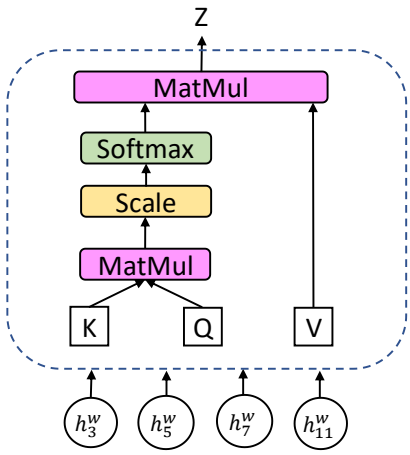

Fig. 2. Illustration of self-attention module.

Table 2. NLI experiment results. Evaluation metric is the classification accuracy in test for SNLI and in matched evaluation for MultiNLI. Result of previous model is come from the paper where the model is proposed if not point out specially.

Models	SNLI	MultiNLI
Bi-LSTM+Maxpooling [5]	84.5	68.2 [15]
shortcut-stacked BiLSTM [16]	86.1	-
DiSAN [17]	85.6	-
Reinforced self-attention [18]	86.3 [19]	-
TC-RN [19]	85.7	-
SPINN [9]	82.6	68.2 [15]
gumbel Tree-LSTM [11]	86.0	69.5 [15]
PLGLH	**86.4**	**72.6**
PLGLH\global	84.3	68.5
PLGLH\local	86.1	71.4

3. EXPERIMENTS AND ANALYSIS

In this section, we conduct a plethora of experiments to study the effectiveness of our model, including several datasets in classification and language inference. In order to study the influence of each module, we conduct module analysis experiments by deleting global encoding module or local encoding module. PLGLH\global refers to the model without global encoding, PLGLH\local refers to the model without local encoding. In all of our experiments, 300-dimensional GloVe [4] word embeddings are used to represent words, out-of-vocabulary words are mapped to zero vectors with same size like others. All the word embeddings retain fixed during training. Minibatch size is set to 64. To train the model, Adam optimizer with learning rate 1e-3 is used. The drop out rate is 0.2.

3.1. Language Inference Task

Natural language inference(NLI) is a fundamental task in the field of NLP that involves reasoning about semantic relationship between sentences. So that the performance of NLI is a very important measurement for sentence embedding quality.

In this experiment, we choose two datasets including Stanford Natural Language Inference(SNLI) dataset [3] and Multi-Genre Natural Language Inference (MultiNLI) dataset [20]. SNLI dataset has 550k human-annotated sentence pairs, called premise and hypothesis, each labeled with one of the following pre-defined relationship: $Entailment$, means the premise entails the hypothesis; $Contradiction$, means they contradict with each other; $Neutral$, means they are irrelevant. MultiNLI is almost same form like SNLI, but covers a range of genres of spoken and written text and supports cross-genre evaluation. There is some examples of them in Table 1.

Following the previous work, we use a siamese architecture that applying the model to both premise and hypothesis. emb_p and emb_h are fixed-length vector representations for premise and hypothesis respectively. The final sentence-pair representation is formed by concatenating the original vectors with the absolute difference and element-wise multiplication between them:

$$emb^{inp} = \left[emb^p; emb^h; \left|emb^p - emb^h\right|; emb^p \odot emb^h\right]$$

At last, we feed the sentence-pair representation emb^{inp} into

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.

a classification network with two layers of linear transformation. The hidden dimension of linear transformation is 512. This is the standard scheme for sentence encoders trained on SNLI.

Experiment results are listed in Table 2. Evaluation metric is the classification accuracy in test for SNLI and in mismatched evaluation set [1] for MultiSNLI. Result in SNLI shows that our model outperform the sequence models like the BiLSTM with maxpooling [5] and shortcut-stacked BiLSTM [16], attention based model like DiSAN [17], get better result than Reinforced-selfattention [18], which use REINFORCE algorithm. Comparement with other latent tree model like SPINN [9] and gumbel Tree-LSTM [11] or model use tree structure as constrain like TC-RN [19], our model also give a better accuracy. When comes to the MultiNLI dataset, there is not so much works about evaluation on this single dataset, we just compare with the baseline model BiLSTM and some latent tree model like SPINN and gumbel Tree-LSTM. Results show that our model can still get better performance.

3.2. Classification Task

Sentence representation is widely applied in single sentence classification, and it is the most common task of NLP. We evaluate our model on various datasets as listed bellow. Table 1 also give some examples of them.

- The Customer Reviews(CR) dataset. Dataset consist of reviews for some products, where the task is to classify each customer review as a positive or negative review [21].
- The Movie Reviews(MR) dataset. Dataset consist of reviews sentence for a movie with an assigned positive or negative label about reviewers attitude [1].
- The Subjectivity(SUBJ) dataset is to classify a sentence as being subjective or objective [22].

For each experiment, we apply our model to get the sentence embedding emb, and then use a classification network with one layer of linear transformation. The hidden dimension of linear transformation is 100.

Experiment results are listed in Table 3. Our model outperform models such as baseline model BiLSTM and CNN-Ana [23], tree constrained models like DC-TreeLSTM [24] and SATA Tree-LSTM [26] or latent tree model HS-LSTM [10] and gumbel Tree-LSTM [11]. Only TreeNet [8] get better result than ours in one task.

But our model requires significantly less time to train. In order to explain intuitively, we record time consuming of other two model: TreeNet, which gets best result but can't compute in batches; gumbel Tree-LSTM, which also apply ST gumbel-softmax to latent tree model, but suffer from heavy calculation for candidate parent tree nodes. Result shows in Fig 3. Result shows our model is at least 10 times faster than existing state-of-the-art method at the training stage. Note that we doesn't compare time consuming with reinforce learning method since it obviously slow and hard to convergence [13].

[1]To see more detail about MultiNLI in `https://www.nyu.edu/projects/bowman/multinli/`

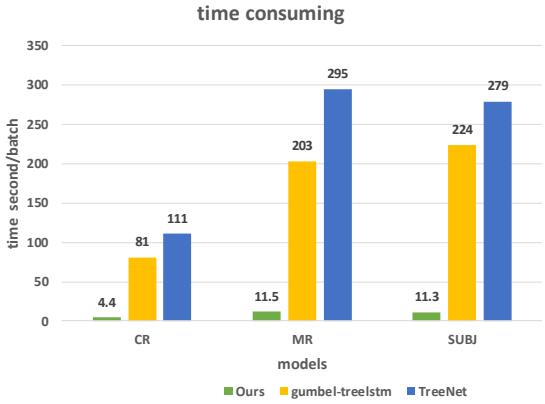

Fig. 3. Comparison of time consumption between our model , gumbel Tree-LSTM and TreeNet. Evaluation metric is training time per batch.

Table 3. Classification experiment result. Comparison of accuracy on test set, result of previous model is come from the paper where the model is proposed except gumbel Tree-LSTM conducted by ourself.

Models	MR	CR	SUBJ
BiLSTM [5]	80.7	83.8	93.4
CNN-Ana [23]	81.0	84.65	93.67
DC-TreeLSTM [24]	81.7	-	93.7
AdaSent [25]	83.1	86.3	95.5
TreeNet [8]	83.8	**88.4**	**95.9**
gumbel Tree-LSTM [11]	83.0	86.1	94.9
HS-LSTM [10]	82.1	-	93.7
SATA Tree-LSTM [26]	83.8	-	95.4
PLGLH	**84.6**	87.4	**95.9**
PLGLH\global	83.7	86.5	95.1
PLGLH\local	83.0	86.0	94.6

3.3. Module Analysis

Results of module analysis experiments show that performance gets slight decrease in all dataset without local encoding or global encoding module. This phenomenon suggests the significance of combining of global and local encoding. Not only that, result of PLGLH\local is better than

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.

PLGLH\global in the natural languague inference experiments, when opposite results in classification tasks. It verifies our suppose since natural languague inference is determined by whole sentence semantic while sentence classification takes more care of specific parts like verb.

4. CONCLUSION

This paper proposes a PLGLH model for sentence embedding which inherits the advantages of existing latent structure models while requires less time complexity and can combine global context dependency information and task-specific local information at phrase level.

From experiment results on a wide range of datasets, we validate the effectiveness of our model. Results empirically show that hierarchical information and global-local is important for understanding sentence and latent tree model is an effective and efficient way to replace recursive tree model.

5. REFERENCES

[1] Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on ACL*. ACL, 2005, pp. 115–124.

[2] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[3] Samuel R Bowman, Gabor Angeli, et al., "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[4] Jeffrey Pennington, Richard Socher, et al., "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on EMNLP*, 2014.

[5] Alexis Conneau, Douwe Kiela, et al., "Supervised learning of universal sentence representations from natural language inference data," *CoRR*, 2017.

[6] Zhouhan Lin, Minwei Feng, et al., "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[7] Kai Sheng Tai, Richard Socher, et al., "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[8] Zhou Cheng, Chun Yuan, et al., "Treenet: Learning sentence representations with unconstrained tree structure.," in *IJCAI*, 2018, pp. 4005–4011.

[9] Samuel R Bowman, Jon Gauthier, et al., "A fast unified model for parsing and sentence understanding," *arXiv preprint arXiv:1603.06021*, 2016.

[10] Tianyang Zhang, Minlie Huang, et al., "Learning structured representation for text classification via reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] Jihun Choi, Kang Min Yoo, et al., "Learning to compose task-specific tree structures," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Haoyue Shi, Hao Zhou, et al., "On tree-based neural sentence modeling," *arXiv preprint arXiv:1808.09644*, 2018.

[13] Eric Jang, Shixiang Gu, et al., "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[14] Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

[15] Adina Williams, Andrew Drozdov*, et al., "Do latent tree learning models identify meaningful structure in sentences?," *Transactions of the ACL*, vol. 6, 2018.

[16] Yixin Nie and Mohit Bansal, "Shortcut-stacked sentence encoders for multi-domain inference," *arXiv preprint arXiv:1708.02312*, 2017.

[17] Tao Shen, Tianyi Zhou, et al., "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] Tao Shen, Tianyi Zhou, et al., "Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling," *arXiv preprint arXiv:1801.10296*, 2018.

[19] Lei Yu, Cyprien de Masson d'Autume, et al., "Sentence encoding with tree-constrained relation networks," *arXiv preprint arXiv:1811.10475*, 2018.

[20] Adina Williams, Nikita Nangia, et al., "A broad-coverage challenge corpus for sentence understanding through inference," *arXiv preprint arXiv:1704.05426*, 2017.

[21] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[22] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on ACL*. ACL, 2004, p. 271.

[23] Ye Zhang and Byron Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[24] Pengfei Liu, Xipeng Qiu, et al., "Dynamic compositional neural networks over tree structure," *arXiv preprint arXiv:1705.04153*, 2017.

[25] Han Zhao, Zhengdong Lu, et al., "Self-adaptive hierarchical sentence model," in *Twenty-Fourth IJCAI*, 2015.

[26] Taeuk Kim, Jihun Choi, et al., "Dynamic compositionality in recursive neural networks with structure-aware tag representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on June 30,2021 at 09:23:59 UTC from IEEE Xplore. Restrictions apply.