CONNECTING TARGETS VIA LATENT TOPICS AND CONTRASTIVE LEARNING: A UNIFIED FRAMEWORK FOR ROBUST ZERO-SHOT AND FEW-SHOT STANCE DETECTION

Rui Liu[1,2], *Zheng Lin*[1,2,*], *Peng Fu*[1,*], *Yuanxin Liu*[1,2], *Weiping Wang*[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{liurui1995,linzheng,fupeng,liuyuanxin,wangweiping}@iie.ac.cn

ABSTRACT

Zero-shot and few-shot stance detection (ZFSD) aims to automatically identify the users' stance toward a wide range of continuously emerging targets without or with limited labeled data. Previous works on in-target and cross-target stance detection typically focus on extremely limited targets, which is not applicable to the zero-shot and few-shot scenarios. Additionally, existing ZFSD models are not good at modeling the relationship between seen and unseen targets. In this paper, we propose a unified end-to-end framework with a discrete latent topic variable that implicitly establishes the connections between targets. Moreover, we apply supervised contrastive learning to enhance the generalization ability of the model. Comprehensive experiments on the ZFSD task verify the effectiveness and superiority of our proposed method.

Index Terms— Stance Detection, Latent Topic Variable, Contrastive Learning, BERT

1. INTRODUCTION

In recent years, with the explosive growth of information on the Internet, it is crucial to identify the underlying stance of a user's utterance, which has triggered a great deal of work on stance detection [1, 2, 3]. Stance is defined as the expression of the speaker's standpoint and judgment toward a given proposition. Previous methods have achieved promising performance in both in-target stance detection [4, 5], in which the model is trained and tested on the same targets, and cross-target stance detection [2, 6, 7], in which the destination targets are different from those in the training set. In terms of in-target stance detection, it requires collecting labeled training data for every new target, which is time-consuming and labor-intensive. In terms of the cross-target setup, while no annotated data is needed for the destination target, it is assumed that the destination target should be strongly related to the training targets by human judgment (i.e., training on a source target and testing on a closly related destination target). However, both setups only considered a very small number of targets, which is oversimplified for real-world application.

In this study, we explore a more practical setting of zero-shot and few-shot stance detection (ZFSD) with a new dataset (i.e., VAST) introduced by Allaway and McKeown [8]. Compared to the in-target or cross-target stance datasets [1, 9, 10] with very limited targets (e,g, SemEval2016 with 6 targets and WT-WT with 5 targets), there are a large number of targets covering extensive topics in the VAST. In the zero-shot scenario, the model is tested on numerous unseen targets without training data. And, in the few-shot scenario, the model is evaluated on those seen targets appearing in the training set but having very few training examples. Therefore, a key challenge for ZFSD is how to model the relationship between vast and multifarious targets without or with limited labeled samples.

Recently, some cross-target models [11, 12] attempted to utilize adversarial training technique to learn the association between targets. However, these models require a large set of unlabled data from the destination target, which is obviously infeasible for the ZFSD. For ZFSD, CKE-Net [13] brought in the commonsense knowledge from the external knowledge base ConceptNet [14] to enhance the connection between the document and the target, but they ignored the relationship among the targets. Allaway and McKeown [8] tackled this problem with a two-stage method (i.e., TGA-Net). First, they cluster the training data based on the original BERT representations without fine-tuning. Then, for each data, the centroid of the nearest cluster is used to guide the prediction. But, a critical problem of TGA-Net is that there is an apparent mismatch between the representations used in two stages if BERT is fine-tuned in the second stage. This is also the reason why their BERT model does not perform further fine-tuning but is fixed, which will greatly limit the performance of the model.

To address aforementioned issues, we propose a unified end-to-end framework for ZFSD with **D**iscrete Latent **T**opic

*Zheng Lin and Peng Fu are the corresponding authors.

This work was supported by the National Natural Science Foundation of China under Grants 61976207 and 61906187.

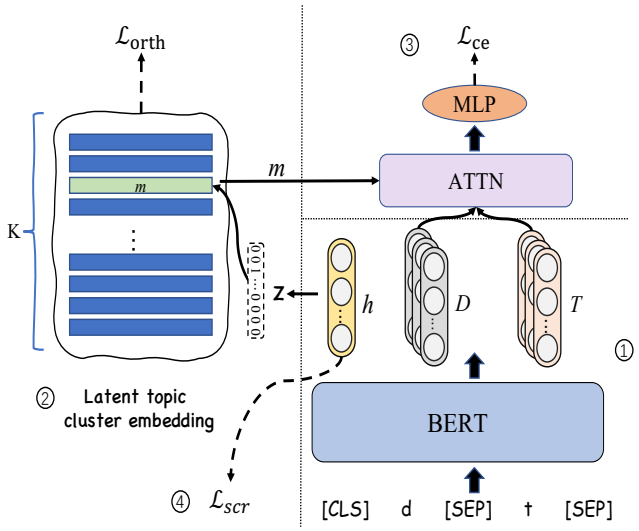

Fig. 1. Overview of our unified end-to-end framework DTCL, which consists of four major modules: (1) a BERT-encoder (§ 2.2), (2) a latent topic inference network (§ 2.3), (3) a stance classifier (§ 2.4), and (4) a supervised contrastive regularization (§ 2.5).

Variable and Supervised **C**ontrastive **L**earning (DTCL), aiming to model the potential relationship between targets. Specifically, we introduce a latent embedding space reflecting the target connections, in which each latent embedding vector represents a latent topic cluster. These embeddings can be learned automatically from the corpus. On this basis, we define a discrete latent topic variable for the targets to correspond to the topic cluster they belong to. In addition, we utilize an orthogonal regularization term to prevent the latent topic clusters from being homogenous. However, we find that the latent topic variable pays more attention to the seen targets. To alleviate this issue, we employ a supervised contrastive regularization to enhance the generalization ability of the model, which can improve the performance in zero-shot and few-shot scenarios simultaneously. Supervised contrastive learning [15, 16] can narrow the gap between examples that belong to the same class but different targets, and push the examples of different classes away from each other. Our model achieves state-of-the-art performance on the VAST, and further experiments demonstrate the effectiveness of our method.

2. METHODOLOGY

2.1. Problem Formulation

Given the zero-shot and few-shot stance detection dataset $D = \{(d_i, t_i, y_i)\}_{i=1}^{L}$, where d_i denotes the input document, t_i denotes the corresponding target, and $y_i \in \{\text{pro, con, neutral}\}$ is the stance label of the document-target pair (d_i, t_i), we aim at developing a classifier to predict the stance label y of each document-target pair (d, t), which can perform well on the zero-shot and few-shot scenarios.

2.2. BERT Encoder

As shown in Figure 1, we first use the pre-trained language model BERT [17] to encode the document $d = \{d^1, \cdots, d^{|d|}\}$ and the target $t = \{t^1, \cdots, t^{|t|}\}$, where $|d|$ and $|t|$ is the number of the document and the target respectively. Specifically, the document d and the target t is concatenated together with special tokens as follows: [CLS] d [SEP] t [SEP]. Then, we extract the output vectors at final layer $\boldsymbol{D} = \{\boldsymbol{d}^1, \cdots, \boldsymbol{d}^{|d|}\}$ for the document and $\boldsymbol{T} = \{\boldsymbol{t}^1, \cdots, \boldsymbol{t}^{|t|}\}$ for the target. Two representation sequences participate in subsequent calculations.

2.3. Latent Topic Inference Network

We denote $\boldsymbol{M} \in \mathbb{R}^{K \times D}$ as the latent topic cluster embedding space, where K is the number of different latent topic vectors, and D is the dimensionality of each latent topic cluster vector. Given the document-target pair, the goal of the latent topic inference network is to obtain a latent topic cluster vector from the embedding space. Latent topic variable $\boldsymbol{z}$ is denoted as a categorical variable following a categorical distribution with K class probabilities $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_K\}$, in which each value corresponds to a particular latent topic cluster.

After obtaining the final hidden states from the BERT encoder, the embedding $\boldsymbol{h}$ of the [CLS] token, which contains the semantic information of the whole sequence, is fed into a linear projection for the class probability $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} = softmax(\boldsymbol{W}_1\boldsymbol{h} + \boldsymbol{b}_1) \in \mathbb{R}^K. \tag{1}$$

Without loss of generality, we assume the latent topic sample z sampled from the categorical distribution with class probability $\boldsymbol{\pi}$ is a K-dimensional one-hot vector. This poses difficulty in practice due to the discrete nature of $\boldsymbol{z}$, which cannot be directly optimized by gradient descent. To address this problem, we use Gumbel-Softmax [18] to approximate the categorical sampling procedure. The approximation $\boldsymbol{p}$ to the one-hot vector z is following as:

$$p_i = \frac{\exp\left(\left(\log\left(\pi_i\right) + g_i\right)/\psi\right)}{\sum_{j=1}^{k}\exp\left(\left(\log\left(\pi_j\right) + g_j\right)/\psi\right)} \quad \text{for } i = 1, \ldots, K, \tag{2}$$

where $g_i = -log(-log(u_i))$, $u_i \sim Uniform(0, 1)$ and ψ is the Gumbel softmax temperature. Eventually, we can select a latent topic cluster embedding $\boldsymbol{m}$ from the latent topic embedding space:

$$\boldsymbol{m} = \boldsymbol{M}^T\boldsymbol{p} \in \mathbb{R}^D. \tag{3}$$

During the test phase, the argmax function is employed on the class probability $\boldsymbol{\pi}$ to create a one-hot vector to select

the corresponding latent topic cluster. For more details about the gumbel softmax trick, please refer to [18, 19]. Moreover, inspired by [20], we use an orthogonal regularization to encourage the latent topic embeddings to be different from each other, which can prevent them from being homogenous. The orthogonal regularization loss is defined as:

$$\mathcal{L}_{orth} = \left\| \boldsymbol{I}_K - \boldsymbol{M}^\top \boldsymbol{M} \right\|_F^2, \tag{4}$$

where $\boldsymbol{I}_K$ is an identity matrix, $\boldsymbol{M}$ is the latent topic embeddings, $\|*\|_F^2$ is the squared Frobenius norm of a matrix.

2.4. Stance Classifier

To establish the deep relationship between the current target and the targets in the same cluster, we employ the scaled dot-product attention function [21] to congregate the information for $\boldsymbol{D}$ and $\boldsymbol{T}$ with the latent topic cluster vector $\boldsymbol{m}$ as query:

$$\boldsymbol{c}_d = \text{ATTN}(\boldsymbol{m}, \boldsymbol{D}, \boldsymbol{D}), \tag{5}$$

$$\boldsymbol{c}_t = \text{ATTN}(\boldsymbol{m}, \boldsymbol{T}, \boldsymbol{T}), \tag{6}$$

in which the three inputs of the attention layer are query, key, and value from left to right. We feed $\boldsymbol{c}_d$ and $\boldsymbol{c}_t$ into multi-layer perceptron to compute the stance probabilities as:

$$\hat{\boldsymbol{y}} = \text{softmax}(\text{MLP}(\boldsymbol{c}_d, \boldsymbol{c}_t)). \tag{7}$$

The cross-entropy loss utilized to train the model is calculated by:

$$\mathcal{L}_{ce} = -\frac{1}{L} \sum_{i=1}^{L} \sum_{j=1}^{C} y_i^j log(\hat{y}_i^j), \tag{8}$$

where $\boldsymbol{y}_i$ is the one-hot vector denoting the stance of the i-th document-target pair (d_i, t_i), and C is the size of the stance label set.

2.5. Supervised Contrastive Regularization

To make the model more robust, supervised contrastive learning is applied as a regularization term to push the data from the different targets but belong to the same class close and the data from different classes further apart, which can learn enhanced target-invariant features. The supervised contrastive regularization loss is formulated as:

$$\mathcal{L}_{scr} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j \in \mathcal{N}_i} \log \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_k / \tau)}, \tag{9}$$

where N is the batch size, $\mathcal{N}_i = \{\boldsymbol{h}_j | i \neq j, y_i = y_j\}$ is the positive examples of $\boldsymbol{h}_i$, N_{y_i} is the number of examples which is labeld as y_i in the same batch, and τ is temperature parameter. The total loss can be defined as:

$$\mathcal{L} = \lambda_c \mathcal{L}_{ce} + \lambda_s \mathcal{L}_{scr} + \lambda_o \mathcal{L}_{orth}, \tag{10}$$

where λ_c, λ_s, and λ_o are scalar weighting hyperparameters.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

In order to verify the effectiveness of our proposed DTCL, we evaluate it on the VAried Stance Topics (VAST) dataset, which consists of thousands of targets. VAST is split according to documents, i.e., documents in the three subsets have no intersection. Table 1 shows the statistics of the dataset. Following [8], the macro-averaged F1 is computed as the evaluation metric.

Table 1. Statistics of VAST dataset.

	Train	**Dev**	**Test**
# Examples	13477	2062	3006
# Documents	1845	682	786
# Distinct Targets	4641	497	759
# Zero-shot Targets	4003	383	600
# Few-shot Targets	638	114	159

3.2. Baselines

We compare our model with two groups of baseline methods. First, we consider in-target and cross-target models based on BiLSTM, such as BiCond [2], CrossNet [6], and SEKT [22], which do not rely on the labeled and unlabeled data of the destination target. Then, we also compare our model with existing ZFSD models, including BERT-joint [8], TGA-Net [8], and CKE-Net [13]. Note that, the BERT-joint can be seen as the variant of the CrossNet. In BERT-joint and TGA-Net, the BERT encoder is fixed during training, which may limit the power of BERT. Therefore, we also re-implement these two methods under fine-tuning.

3.3. Implementation Details

Following [8], the maximum sequence lengths of the document and the target are 200 and 5 respectively. We use the base version of BERT as the encoder. We train our model for 10 epochs, using the Adam [23] optimizer with a learning rate of 3e-5. The batch size is set to 64. A linear learning rate scheduler is applied with linear warm-up steps 0.1 of total training steps. The hyperparameter combination for the total loss is $\lambda_c = 0.1$, $\lambda_s = 0.9$, and $\lambda_o = 0.0001$. The number of latent topic clusters K is 75. The Gumbel softmax temperature parameter ψ is set to 0.5, and the temperature parameter τ for supervised contrastive regularization is set to 0.1.

3.4. Results and Discussions

3.4.1. Results of Different Scenarios

Table 2 shows the results of our model and several state-of-the-art baselines on both zero-shot and few-shot scenarios. As

Table 2. F1 (%) on the test set. * represents our implementation, and the suffix "ft" means the bert is fine-tuned during training.

Model	F1 Zero-Shot				F1 Few-Shot				F1 All			
	pro	con	neu	all	pro	con	neu	all	pro	con	neu	all
BiCond [2]	45.9	47.5	34.9	42.7	45.4	46.3	25.9	39.2	45.7	46.8	30.6	41.0
Cross-Net [6]	46.2	43.4	40.4	43.4	50.8	50.5	41.0	47.4	48.6	47.1	40.8	45.5
SEKT [22]	50.4	44.2	30.8	41.8	51.0	47.9	21.5	47.4	50.7	46.2	26.3	41.1
BERT-joint [8]	54.6	58.4	85.3	66.0	54.3	59.7	79.6	64.6	54.5	59.1	82.3	65.3
TGA-Net [8]	55.4	58.5	85.8	66.6	58.9	59.5	80.5	66.3	57.3	59.0	83.1	66.5
BERT-joint-ft*	57.9	60.3	87.5	68.5	59.5	62.1	83.1	68.4	58.8	61.4	85.3	68.4
TGA-Net-ft*	56.8	59.8	**88.5**	68.4	62.8	60.1	83.4	68.7	59.9	59.9	85.9	68.6
CKE-Net [13]	**61.2**	61.2	88.0	70.2	**64.4**	62.2	83.5	70.1	**62.9**	61.7	85.7	70.1
DTCL	60.0	**64.7**	87.6	**70.8**	63.0	**66.3**	**85.5**	**71.6**	61.6	**65.5**	**86.5**	**71.2**

Table 3. Experimental results (F1 %) of ablation study. Here, "ltv" denotes the latent topic variable, "scr" denotes the supervised contrastive regularization, "orth" means the orthogonal regularization, and "all" means all our proposed components.

Model	Zero-Shot	Few-Shot	ALL
DTCL	**70.8**	**71.6**	**71.2**
w/o ltv	70.1	69.6	69.9
w/o scr	69.3	70.8	70.1
w/o scr&orth	68.9	70.0	69.5
w/o all	68.5	68.4	68.4

we can see, our model DTCL clearly outperforms the baselines in all scenarios. When BERT is fixed, TGA-Net outperforms BERT-joint. However, the comparison between BERT-joint-ft and TGA-Net-ft shows that the advantage of the two-stage strategy (cluster-then-predict) is less obvious under fine-tuning. More importantly, DTCL is remarkably better than TGA-Net-ft, which implies that our end-to-end framework is good at associating the relationship between seen and unseen targets. Comparative results between BERT-based models and conventional LSTM models show that the BERT can capture more valuable information from the ZFSD dataset.

3.4.2. Ablation Study

We carry out a series of ablation experiments to analyze the effect of each component of our method. We can observe from Table 3 that the removal of either the latent topic variable or the supervised contrastive regularization can degrade the model performance drastically. Thus, both two modules are crucial for our model. Moreover, the improvement of using only latent topic variable in the zero-shot scenario is much lower than that in the few-shot scenario, which implies that the latent topic variable maybe focus more on the seen targets. In comparison, the supervised contrastive regularization can contribute in two scenarios almost equally, which indicates that the regularization can promote the model to learn more distinguishable representation and be more robust. Besides, the orthogonal regularization benefits the learning of target clustring.

Table 4. Accuracy on five challenging phenomena in the test set.

Model	Imp	mlT	mlS	Qte	Sarc
BERT-joint	57.1	59.0	52.4	63.4	60.1
TGA-Net	59.4	60.5	53.2	66.1	63.7
BERT-joint-ft	61.7	62.1	54.7	64.7	66.8
TGA-Net-ft	61.5	62.5	54.6	66.4	67.5
CKE-Net	**62.5**	63.4	55.3	69.5	68.2
DTCL	62.1	**63.6**	**55.6**	**69.9**	**70.3**

3.4.3. Results of Challenging Phenomena

Following [8, 13], we evaluate our model on five challenging phenomena in the test set: (1) Imp, the target with neutral label and not contained in the document, (2) mlT, a document has examples with multiple targets,(3) mlS, a document has examples with different, non-neutral, stance labels, (4) Qte, a document with quotations, and (5) Sarc, a document with sarcasm. We observed that under the five challenge phenomena, our method has a great improvement over BERT-joint-ft and TGA-Net-ft, which proves the expressive representation ability of the model. Imp mainly investigates the ability of modeling the relationship between document and target. CKE-Net has natural advantages for modeling document-target relationship based on relational graph and external knowledge base. Without using external knowledge, our method has reached a similar level of reasoning as CKE-Net on Imp.

4. CONCLUSION

In this paper, we propose a unified end-to-end framework for the zero-shot and few-shot stance detection tasks. In particular, we introduce a latent topic cluster embedding and a discrete latent topic variable to build a bridge between various targets. Moreover, we adopt a supervised contrastive regularization to enhance the generalization of the model. The experimental results on the VAST dataset show that our method improves performance significantly and achieves the best results.

5. REFERENCES

[1] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *SemEval-2016*, pp. 31–41.

[2] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva, "Stance detection with bidirectional conditional encoding," in *EMNLP 2016*, pp. 876–885.

[3] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Techn.*, vol. 17, no. 3, pp. 26:1–26:23, 2017.

[4] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui, "Stance classification with target-specific neural attention," in *IJCAI-17*, pp. 3988–3994.

[5] Penghui Wei, Wenji Mao, and Guandan Chen, "A topic-aware reinforced model for weakly supervised stance detection," in *AAAI 2019*, pp. 7249–7256.

[6] Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks, "Cross-target stance classification with self-attention networks," in *ACL 2018*, pp. 778–783.

[7] Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu, "Target-adaptive graph for cross-target stance detection," in *WWW 2021*, pp. 3453–3464.

[8] Emily Allaway and Kathleen McKeown, "Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations," in *EMNLP 2020*.

[9] Jonathan Kobbe, Ioana Hulpus, and Heiner Stuckenschmidt, "Unsupervised stance detection for arguments from consequences," in *EMNLP 2020*, pp. 50–60.

[10] Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier, "Will-they-won't-they: A very large dataset for stance detection on twitter," in *ACL 2020*, pp. 1715–1724.

[11] Penghui Wei and Wenji Mao, "Modeling transferable topics for cross-target stance detection," in *SIGIR 2019*, pp. 1173–1176.

[12] Emily Allaway, Malavika Srikanth, and Kathleen R. McKeown, "Adversarial learning for zero-shot stance detection on social media," in *NAACL-HLT 2021*, pp. 4756–4767.

[13] Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang, "Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph," in *ACL 2021*, 2021, pp. 3152–3157.

[14] Robyn Speer, Joshua Chin, and Catherine Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *AAAI 2017*, pp. 4444–4451.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," in *NeurIPS 2020*.

[16] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *ICLR 2021*.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, pp. 4171–4186.

[18] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR 2017*.

[19] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR 2017*.

[20] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *NeurIPS 2016*, pp. 343–351.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS 2017*, pp. 5998–6008.

[22] Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai, "Enhancing cross-target stance detection with transferable semantic-emotion knowledge," in *ACL 2020*.

[23] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*.