CROSS-TARGET STANCE DETECTION VIA REFINED META-LEARNING

Huishan Ji[1,2], *Zheng Lin*[1,2,*], *Peng Fu*[1,*], *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
{jihuishan,linzheng,fupeng,wangweiping}@iie.ac.cn

ABSTRACT

Cross-target stance detection (CTSD) aims to identify the stance of the text towards a target, where stance annotations are available for (though related but) different targets. Recently, models based on external semantic and emotion knowledge have been proposed for CTSD, achieving promising performance. However, such solutions rely on much external resources and harness only one source target, which is a waste of other available targets. To address the problem above, we propose a many-to-one CTSD model based on meta-learning. To make the most of meta-learning, we further refine it with a balanced and easy-to-hard learning pattern. Specifically, for multiple-target training, we feed the model according to the similarity among targets, and utilize two kinds of re-balanced strategies to deal with the imbalance in data. We conduct experiments on SemEval 2016 task 6, and results demonstrate that our method is effective and establishes a new state-of-the-art macro-f1 score for CTSD.

Index Terms— Meta-Learning, Stance Detection, Curriculum Learning

1. INTRODUCTION

Stance detection aims to identify the attitude of the text towards a topic (referred as target), of which labels can be favor, against, discuss and so on [1]. However, it is neither economically ideal nor technically necessary to collect ample annotated samples from every target, especially when new targets cost time to raise enough opinions from users. Additionally, it is such a waste of data if we leave samples from other targets alone. In a word, there is a need to learn a model that can utilize data from other targets in order to perform better on new targets for stance detection.

Cross-target stance detection, or CTSD, consists of two objects, which are stance detection and transfer learning. For stance detection, besides traditional methods like LSTM or CNN, their variants have been proposed and achieved fair results [2][3]. As records these models has made in their era, the birth of pretrained model, like BERT [4], further improved the performance. For the other object, the transfer learning problem in stance detection, there are two typical ideas, which are extracting common features [5] and importing external knowledge [6][7]. The former focuses on extracting shared features. The latter two either utilizes Semantic-Emotion Knowledge Transfer (SEKT), which introduces external knowledge [6] into the model, or target-adaptive pragmatics dependency graphs (TPDG) [7], which utilizes pragmatics knowledge to analyze stance-related information in both source target and destination target. Effective as they are, SEKT and TPDG are not convenient enough to conduct, given the acquisition of a semantic-emotion heterogeneous graph from external semantic and emotion lexicons could cost extra endeavor. Therefore, though all these models have their peculiarities, there is a common problem here, that they utilize only one source target to help the destination target, which is a waste of other available targets.

Therefore, we consider about harnessing multiple targets. However, such a many-to-one CTSD problem presents challenge. Simply feeding the model with multiple targets barely brings any improvement [8], which is also validated in our comparison experiments. From another prospective, as multiple source targets come from different realms but share the same goal, they can be regarded as different but similar tasks. Such inspiration led us to take meta-learning [9] into account. Meta-learning focuses on generalizing better in new tasks based on existing tasks, which is ideal for such a problem.

In addition, as there are multiple source targets, instead of a typical method randomly feeding the targets into the model [8], we propose two strategies for improvement. For multiple tasks, inspired by curriculum learning [10], we explore a calculation method to create a target sequence, according to which the model trains better. For the imbalance in the dataset, refining strategies are applied to help reshape the data.

Note that our object here is not to design a highly complicated algorithm or model to significantly surpass the best existing performance, but to study applying meta-learning to solve such a many-to-one transfer learning problem in stance

*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.

detection, and to explore the potential of MAML under the circumstance of an imbalanced dataset by utilizing two refining strategies.

2. RELATED WORK

2.1. Stance detection

Stance Detection aims to infer the attitude of a text towards specific target expression. Early stance detection studies were concentrated on debates [12], where multiple studies were based on variants of RNN or CNN [2][3] and achieved fair results. With the birth of attention mechanism, studies applying attention to stance detection have emerged, like a basic hierarchical attention model [13][14] or BERT [4] and its variants [15][16]. Though these methods train well on single target in stance detection, they are not designed for cross-target problems. In order to deal with multiple targets, recently there are studies importing external knowledge in order to bridge the gap between targets [6][7]. The imported knowledge includes semantic and emotion knowledge (SEKT) [6] and pragmatics knowledge (TPDG) [7]. The former utilizes external semantic and emotion dictionaries, which expands the LSTM cell with an additional memory unit, to build a semantic-emotion heterogeneous graph, and applies GCN to learn the representations. The latter focuses on pragmatic features, extracting in-target pragmatic knowledge and cross-target knowledge, exploring the different roles of words in different targets, based on a GCN block.

2.2. Meta-learning

Meta-learning, or learning to learn, treats learning as a parametrized algorithm [9], even referred as a building block in artificial intelligence [17]. Meta-learning tries to teach the model how to learn, and there are three typical approaches: Model-based [18], Metric-based [19] and Optimization-based [20][21]. Model-based methods leverage models to generate parameters instead of labels, to initialize another model for classification [18]. Metric-based methods focus on modelling the distance between samples, shortening distances of samples with same labels and pushing away samples of different labels [19]. The model-agnostic meta-learning (MAML) [20] we used in this paper belongs to optimization-based meta-learning. Its idea is to locate a set of initial parameters that can update well on all tasks sampled from data, and generalize to new tasks.

3. METHODOLOGY

3.1. Problem review

Source targets: given sets of claims divided by different targets, where each claim in each set is paired with a corresponding target and annotated by a label, the *i* th set can be expressed as:

$$D_{s_i} = \{c_{j,s_i}, l_{j,s_i}, t_{s_i}, j \in \{0, 1, ..., N_{s_i}\}\}, i \in \{0, 1, ..., N_s\}$$

where s_i stands for the *i* th source target and N_{s_i} is the amount of samples in s_i. There are N_s source targets in total. c_{j,s_i} represents the claim of sample *j* in source target s_i, and l_{j,s_i}, t_{s_i} are the label and target of sample *j*, respectively. And we have a destination set:

$$D_d = \{\{c_j, l_j, t\}, j \in \{0, 1, ..., N_d\}\}$$

where c_j, l_j and t stand respectively for the claim, label and target of sample j in destination target, and N_d is the number of its samples. Note that the destination target here is not fixed and can be any target in the dataset.

The goal of many-to-one CTSD is to sufficiently harness data in each D_{s_i} and use claim c_j and target *t* in D_d to predict the correct corresponding label l_j.

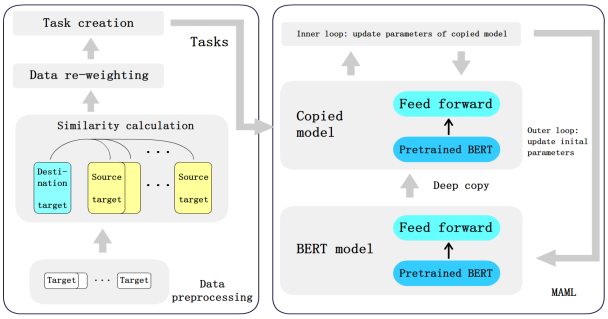

Fig. 1. The structure of the proposed model

3.2. Model description

3.2.1. *Model structure*

Figure 1 shows the overall structure of the model. Our model consists of two main parts: the data preprocessing part and MAML part, as marked at the lower right corners in the figure. The data preprocessing part includes the calculation of similarity among targets, sub-sampling and the creation of tasks, and the MAML part consists of the double loop according to MAML based on BERT. We give detailed description in following sections.

3.2.2. *BERT based MAML*

The training procedure is shown in Algorithm 1 [20] . θ' represents the updated parameters of the copied model, as shown in figure 1. f_θ is the BERT model with initial parameters θ. L is the cross-entropy loss.

The copied model corresponds to f'_θ in algorithm 1, which is updated by samples in the support sets (D). After training on the current batch of tasks, the model is evaluated on the

Algorithm 1 Training procedure

Require: $p(T)$: distribution over tasks
Require: α, β: step size parameters
randomly initialize θ
while not done **do**
  Sample batch of tasks $T_i \sim p(T)$
  for all T_i **do**
    Sample *K* datapoints $D_i = \{x^{(j)}, y^{(j)}\}$ from T_i
    Evaluate $\nabla_\theta L_{T_i}(f_\theta)$ using D_i and L_{T_i} by multi-class cross entropy loss
    Compute adapted BERT parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta L_{T_i}(f_\theta)$
    Sample datapoints $D_i' = \{x^{(j)}, y^{(j)}\}$ from T_i
    Use D_i' to update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i}(f_{\theta_i'})$
  end for
end while

query sets (D') and the gradient is then used for updating the initial parameters θ in the BERT model.

*3.2.3. **Similarity calculation***

We compute the KL divergence based on the distribution of words. Given a source target s_i containing N_{s_i} samples, where each sample consists of a claim c_{s_i}, a stance label l_{s_i}, and a target label t_{s_i}, we process as follows: First, we create vocabulary dictionaries $V_i, i \in 0, 1, ..., N$for all targets, which contains the frequency of each word existed in all the claims of the corresponding target. Second, we calculate the destination target's KL divergence towards each source target:

$$D_{i,KL}(V_d|V_{s_i}) = \sum_j (V_d(j) log \frac{V_d(j)}{V_{s_i}(j)}, i \in \{0, 1, ..., N\}$$

In the calculation above, $D_{i,KL}$ is the destination target V_d's KL divergence towards each source target V_{s_i}, and N is the number of source targets. Third, we rank $D_{i,KL}$ in descending order, and pass the corresponding target sequence into the next module, which is the data re-weighting module.

*3.2.4. **Data re-weighting strategies***

Table 1 shows the label distribution of samples in all targets. Most targets suffer from the imbalance among the numbers of different labels. Due to such condition, we utilize two strategies: sub-sampling and focal loss. For each target, we locate the lowest number of labels M_i, and randomly sample in other two labels, up to M_i, adjusting the numbers of labels to the same in each target.

Assuming y is the real labels and y' is the predicted scores, the focal loss [11] can be computed by:

$$L_f = -(1 - y'y)^\gamma log\, y'y$$

Target	Against	Favor	None	Total
HC	565	163	256	984
LA	545	167	222	934
A	465	124	145	734
CC	27	335	203	565
FM	512	268	170	950
Total	2114	1057	996	4167

Table 1. Label distribution

4. EXPERIMENTS

In this section, we introduce the settings of experiments and the analysis of the results.

4.1. Dataset and evaluation metrics

We conduct experiments on SemEval-2016 Task6. Table 1 shows the statistics of the data. We selected five targets from the dataset, which are Hillary Clinton (HC), Legalization of Abortion (LA), Atheism (A), Climate Change is a Real Concern (CC) and Feminist Movement (FM). As this is a relatively small dataset, we use 8-fold cross validation. In addition, macro f1-score is set as the evaluation criteria.

4.2. Comparison models

In order to validate the effectiveness of our model, we introduce several related models for comparison, including three typical models BiCond, TextCNN-E and BERT, and two state-of-the-art models SEKT and TPDG.

BiCond: Adopting bidirectional conditional encoding to learn both the sentence and the target representation for detecting stance expression [2].

TextCNN-E: A variant of TextCNN for CTSD task [3]. The idea is to integrate semantic and emotional knowledge into each word and expanded the dim of each word vector.

BERT: The well-known model for NLP tasks by Google [4]. Trained on multiple source targets and one destination target.

SEKT: Imports external dictionaries and builds a semantic-emotion graph [6], as introduced in related work.

TPDG: A target-adaptive pragmatics dependency graphs model (TPDG) [7], as introduced in related work.

4.3. Implementation details

We set the maximum sequence length to be 70, which covers all samples. The BERT model we use for comparison and main experiments is bert-base-uncased [4]. We set the size of D_i and D_i' in the algorithm chart to be 40 and 20, which are also known as the sizes of support set and query set. We create 416 tasks for each target. The learning rates α and β are respectively 2e-5 and 3e-5. The batch sizes of outer loop

	HC	LA	A	CC	FM
TextCNN-E[3]	52.6	51.2	52.7	53.6	49.3
Bicond[2]	53.4	63.3	58.2	59.1	55.0
BERT[4]	62.5	64.1	71.3	56.8	61.1
SEKT[7]	72.3	70.0	74.8	-	64.9
TPDG[8]	**75.4**	76.7	76.9	63.0	70.0
Our model	75.0	**77.1**	**83.4**	**65.2**	**72.3**

Table 2. Macro F1-score of CTSD experiments

Model	FM
Complete model	72.3
w/o sub-sampling	67.9
w/o focal loss	71.1
w/o sub-sampling and focal loss	64.8
w/o multiple source targets	63.4

Table 3. Ablation experiments

and inner loop in the algorithm chart are 16 and 8. The update steps in the inner loop is 7.

4.4. Experiment results and analysis

4.4.1. Main experiments

Table 2 shows the results of CTSD task, with best scores marked. Compared with BERT, combining MAML and refining strategies with BERT does contribute a lot to the performance and surpasses almost all models here on all targets but TPDG on *Hillary Clinton*. Note that scores on *CC* are significantly worse than all other targets. As we speculate, the cause to such phenomenon is the extreme imbalance in labels, where only dozens of *Against* are available.

4.4.2. Ablation experiments

Table 3 is the results of ablation experiments, setting FM as the destination target. As shown, sub-sampling contributes a lot to BERT based MAML, and focal loss plays a less important but unignorable role. From another prospective, focal loss and sub-sampling both deal with the imbalance in data, though focal loss brings an improvement of 3.1%, not as effective as sub-sampling (6.3%). Specifically, from row 1 and 3, we can see that when combined with sub-sampling, applying focal loss only brings about a 1.2% increase, far less than that without sub-sampling (shown by row 2 and 4), which means the function of sub-sampling and focal loss partially overlaps. Therefore, either of these two methods can deal a lot with the imbalance in data, but the effect of combining them together is not simply the combination of their individual effects. In addition, the last row in the table shows that training on multiple source targets performs much better than training on the most similar source target (*atheism*) alone, which brings a 8.9% increase in the f1-score.

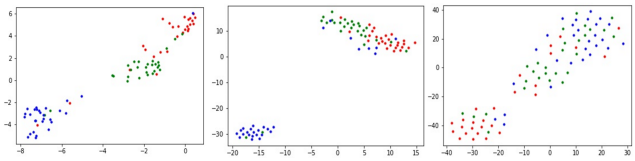

Fig. 2. Scatter diagrams of using sub-sampling (left) or not (middle), and BERT with sub-sampling and focal loss (right). Red, green and blue points stand for the real labels of samples.

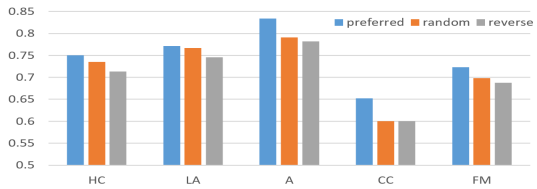

Fig. 3. F1-scores from different feeding orders

To take a deeper look, we drew scatter diagrams based on T-sne to help analyze the results. Figure 2 presents contradiction of scatter diagrams of the trained model on a balanced test set.

As shown in figure 2, it is clear that the model combined with sub-sampling generates the representation much better, as areas of different samples with different labels have less overlapping and possess fewer misplaced samples. Therefore, sub-sampling does contribute a lot.

The diagram on the right is BERT with sub-sampling and MAML is not applied, which is far worse than our complete model on the left. Such phenomenon confirms that when combined with MAML and refining strategies, BERT trains much better, and our strategy applying MAML for BERT works.

Figure 3 shows the results from different feeding orders on all targets. As shown here, all results from random or reverse orders are not as good as the preferred orders, validating the effectiveness of our strategy that feed according to the similarity calculation of targets.

5. CONCLUSION

In this paper, we propose a model for many-to-one CTSD tasks. Our model utilizes multiple source targets to improve the performance on destination target and applies reweighting strategies for the imbalance in data. In particular, our model does not import external knowledge or extract common features among targets, therefore it is more convenient to deploy industrial usage, like in social media. The effectiveness is proven by our experiments and our model shows advancement compared to multiple related models. In addition, we explored the potential of MAML with re-fining strategies, which can be promoted to other problems as well.

6. REFERENCES

[1] I Augenstein, T Rocktaeschel, A Vlachos, K Bontcheva. Stance detection with bidirectional conditional encoding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin. 876–885. 2016.

[2] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, Kalina Bontcheva. Stance Detection with Bidirectional Conditional Encoding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin. 876–885. 2016.

[3] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing . Doha. 1746–1751. 2014.

[4] Iulia Turc, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv:1908.08962v2. 2019.

[5] Penghui Wei Wenji Mao. Modeling Transferable Topics for Cross-Target Stance Detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research Development in Information Retrieval. Paris. 1173–1176. 2019.

[6] Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, Kuai Dai. Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. 3188-3197. 2020.

[7] Bin Liang, Yonghao Fu, Lin Gui, et al. Target-adaptive Graph for Cross-target Stance Detection. In Proceedings of the Web Conference 2021. Association for Computing Machinery. New York. 3453–3464. 2021.

[8] Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, Jianfeng Gao. Multi-task Learning with Sample Reweighting for Machine Reading Comprehension. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis. 2644–2655. 2019.

[9] Sepp Hochreiter, A Steven Younger, Peter R Conwell. Learning to learn using gradient descent. In International Conference on Artificial Neural Networks. Springer. 87–94. 2001.

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, Jason Weston. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning. New York. 41-48. 2009.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice. 2980-2988. 2017.

[12] Marilyn A Walker, Pranav Anand, Robert Abbott, Ricky Grant. Stance classification using dialogic properties of persuasion. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics. Montreal. 592–596. 2012.

[13] Jiachen Du, Ruifeng Xu, Yulan He, Lin Gui. Stance classification with target-specific neural attention networks. In Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne. 3988–3994. 2017.

[14] Qingying Sun, Zhongqing Wang, Qiaoming Zhu, Guodong Zhou. Stance Detection with Hierarchical Attention Network. In Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe. 2399-2409. 2018.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942.

[17] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman. Building machines that learn think like people. arXiv: 1604.00289.

[18] Tsendsuren Munkhdalai Hong Yu. Meta networks. In Proceedings of the 34th International Conference on Machine Learning. Sydney. 2554–2563. 2017.

[19] Jake Snell, Kevin Swersky, Richard Zemel. Prototypical networks for few-shot learning. In Advances in 2017 Neural Information Processing Systems. Long Beach. 4077–4087. 2017.

[20] Chelsea Finn, Pieter Abbeel, Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning. Sydney.1126-1135. 2017.

[21] Alex Nichol John Schulman. 2018. Reptile: a scalable metalearning algorithm. arXiv:1803.02999.

[22] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, Jason Weston. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning. New York. 41–48. 2009.