A Meta-Learning Approach for Few-Shot (Dis)Agreement Identification in Online Discussions

Anonymous ACL submission

Abstract

Online discussions are abundant with different opinions for a common topic, and identifying agreement and disagreement between online posts enables many opinion mining applications. Realizing the increasing needs to analyze opinions for emergent new topics (e.g., from "mask mandate" to "COVID vaccination") that however tend to lack annotations, we present the first meta-learning approach for *few-shot* (dis)agreement identification on a new topic with few labeled instances. We further design a lexicon based regularization loss and propose domain-aware task augmentation for meta-training to enable the meta-learner to learn both domain-invariant cues and domain-specific expressions for (dis)agreement identification. Extensive experiments on two benchmark datasets and evaluation on three topic domains demonstrate the effectiveness of the meta-learning approach that consistently and noticeably outperforms the conventional transfer learning approach based on fine-tuning.

1 Introduction

As seen in many online forums and Subreddits, people express different opinions and perspectives toward a common topic in online discussions. Detecting agreement and disagreement relations between online posts addressing a shared topic will enable many opinion mining applications and inform policy making. However, realizing that new topics keep emerging (e.g., from "mask mandate" to "COVID vaccination"), it is unrealistic to expect existing annotated datasets to cover each topic of interest. To avoid the time-consuming process to create a large annotated dataset for a new topic, we study *few-shot* agreement and disagreement identification that aims to quickly build a model on a new topic domain with few labeled instances.

The traditional transfer learning approach that trains a model on annotation-rich domains and then fine-tunes the model for a new domain usually suffers from the overfitting problem (Zhang et al., 2017) when the number of labeled samples in the target domain is very small. It is known that fine-tuning on a limited number of samples often leads the model to simply memorize the labels for these samples and fail to learn generalizable features for the new domain.

To tackle the difficulty of *few-shot* (dis)-agreement identification under a new topic domain, we present a metric-based meta-learning approach that trains a meta-learner on annotation-rich domains and adapts the meta-learner to a new domain with very few labeled instances. The meta-learner takes K labeled samples per class in the target new domain as support set, attentively builds class embeddings using the support set, and compares a test instance with each class embedding via a learned relation network (Sung et al., 2018) to make a prediction. To mimic the meta-testing procedure and make the model accustomed to the few-shot environment, the meta-training process adopts episodic training (Vinyals et al., 2016) to train the meta-learner: in each training episode, we sampled instances from training domains, including K examples per class as support set and a query set as well, compared each query instance with class embeddings derived from the support set and minimized the loss on the query set.

Inspired by prior research (Misra and Walker, 2013) that studied rich domain-independent indicators of agreement and denial in online discussions, we further encourage the meta-learning system to learn domain-invariant features and thus enhance its ability of quickly generalizing to a new test domain. Specifically, guided by (Misra and Walker, 2013), we compiled a lexicon of domain-independent (dis)agreement indicators consisting of several hundred words and short phrases, e.g., "yes", "make sense", "no" and "but". Then, we designed a *regularization loss* based on the lexicon and added it to the meta-learning system so that the meta-learner pay more attention to the domain-independent cues.

Meanwhile, we decompose an entire training dataset to clusters for episodic training, with each cluster corresponding to a topic domain, in order to better train the meta-learner to recognize domain-specific expressions of agreement and disagreements. Existing labeled datasets for agreement and disagreement identification usually contain data instances from multiple topic domains. If we randomly sample from an entire dataset for each meta-training episode, many episodes have sampled instances in the support set and the query set that do not match in domain and have divergent data distributions. The domain mismatch will lead to poor transfer between support and query sets (Murty et al., 2021), and thus hinders the meta-learner from learning to recognize domain-specific expressions of agreement and disagreements. Therefore, we perform domain-aware task augmentation for meta-training to strengthen the few-shot adaptation ability of the meta-learner, where we sample instances from the same domain to form support set and query set for each episode.

We experiment on two benchmark datasets for agreement and disagreement identification, the Internet Argument Corpus (IAC) (Walker et al., 2012) and the Agreement by Create Debaters corpus (ABCD) (Rosenthal and McKeown, 2015). Evaluation on three topic domains shows that compared to the conventional transfer learning, the meta-learning approach achieves consistent and noticeable performance gains across the three domains under the challenging *few-shot* setting for (dis)agreement identification. Both of the two strategies for strengthening the adaptation ability of the meta-learner further improve the performance of the meta-learner, by enabling it to learn both domain-invariant cues and domain-specific expressions for (dis)agreement identification.

To summarize, our contributions are mainly three:

- We present the first meta-learning approach for *few-shot* agreement and disagreement identification.
- We designed a lexicon based regularization loss to encourage the meta-learner to learn domain-invariant features.
- We perform domain-aware task augmentation for meta-training to better train the meta-learner to recognize domain-specific expressions of agreement and disagreements.

2 Related Work

Research on agreement and disagreement detection in online conversations or social media dialogues attracted increasing attentions. (Walker et al., 2012) provided the Internet Argument Corpus (IAC), annotating agreement/disagreement relation for Q-R (Quote-Response) post pairs in ten different domains, where Response is a single post replying to the previous post Quote. (Misra and Walker, 2013) conducted binary classification (*agreement vs. disagreement*) on the IAC corpus and studied rich domain independent cues for (dis)agreement identification. (Wang and Cardie, 2016) proposed to improve three-way classification (*agreement vs. neutral vs. disagreement*) with a socially-tuned sentiment lexicon. (Rosenthal and McKeown, 2015) introduced a larger dataset, the Agreement by Create Debaters (ABCD) corpus, and conducted three-way classification with transfer learning. However, none of the prior research has studied the (dis)agreement identification task under the cross-domain few-shot setting.

Meta-learning has been studied for years as a general method for few-shot learning. Metric-based meta-learning learns a distance function between data instances and classifies test instances by comparing them to K labeled samples. Several metric-based meta-learners have been proposed, including Siamese Network (Koch et al., 2015), Matching Network (Vinyals et al., 2016), Prototype Network (Snell et al., 2017) and Relation Network (Sung et al., 2018), which learn an embedding function mapping individual instances into a representation space and learn a similarity function to calculate distance between two instances. Another direction is optimization-based meta-learning (Finn et al., 2017) that aims to learn a good initialization to make a neural model reach the optimal for a new task quickly. We focus on developing a metric-based meta-learning model on the basis of Prototype and Relation Network models.

Meta-learning has been used for many NLP tasks under the few-shot setting, including topic classification (Jiang et al., 2018), entity relation classification (Sun et al., 2019; Geng et al., 2019), word sense disambiguation (Deng et al., 2020) and event detection (Deng et al., 2020; Lai et al., 2020). Mostly, prior works used meta-learning to identify unseen new classes and treat a class as a task, we, however, aim to identify (dis)agreement in unseen new domains and treat a domain as a task.

Domain generalization has been studied long before the emergence of meta-learning, aiming to generalize from a set of seen domains to unseen domains without accessing any instance from the unseen domain during the training stage. As a strategy to achieve domain generalization, (Blanchard et al., 2011; Li et al., 2018; Muandet et al., 2013) proposed extracting domain-invariant features from various seen domains to enhance generalization ability. To the best of our knowledge, we lead on using domain-invariant features together with meta-learning to enhance few-shot generalization ability across domains.

Lack of well-defined data distribution is a recognized obstacle of meta-learning for solving NLP problems, generating some attempts in augmenting meta-training tasks. Task augmentation for meta-learner was first studied in (Rajendran et al., 2020). (Bansal et al., 2020) proposed the SMLMT method to create new self-supervised tasks. Most closely related to our work is the strategy mentioned in (Murty et al., 2021) which clustered the entire dataset into several clusters by K-means and sampled support & query set from the same cluster to form training tasks. Our idea of task augmentation is different from theirs in that we relied on domain information to decompose the entire dataset into different training domains, creating clearer boundaries for different types of tasks.

3 The Meta-Learning Approach

In this section, we will elaborate our meta-learning approach in details. Firstly, we introduced the structure of the basic meta-learning model. Then we enhanced the model's domain generalization ability from two perspectives: (1) Manually created a lexicon for *domain-independent* (dis)agreement indicators, and designed a regularization loss to make the meta-learner focus more on domain-invariant features, (2) Decomposed the entire training dataset into several sub-datasets based on *domain-specific* info to augment the task distribution. Fig. 1 illustrated the pipeline of our meta-learning approach.

3.1 The Basic Meta-Learning Model

In the cross-domain few-shot (dis)agreement identification problem, we are given a training dataset $\mathcal{D}_{meta-train}$ consisting of rich labeled Q-R pairs from various domains, and a testing dataset $\mathcal{D}_{meta-test}$ in an unseen new domain. $\mathcal{D}_{meta-test}$ is splitted into two parts: a support set $\mathcal{D}_{test-support}$ with only a small number of K labeled Q-R pairs per class, and a test set $\mathcal{D}_{test-query}$ used to evaluate the model performance on. Our goal is to train a meta-learner $f : (S, x) \rightarrow \hat{y}$ that takes a support set $S = \{s_k^i, i \in 1 \ldots C, k \in 1 \ldots K\}$ and a test instance x as input, then returns a prediction $\hat{y}$ for the instance (x, y), where $y \in \{1, \ldots, C\}$ is the true label, C is the number of classes. The few-shot problem is often named a C-way K-shot learning problem.

3.1.1 Episodic Training

To mimic the meta-testing task that takes a support set $\mathcal{D}_{test-support}$ & test instances $\mathcal{D}_{test-query}$ as input, and make the model accustomed to the few-shot environment, we followed the episodic training idea in (Vinyals et al., 2016) to create training tasks: randomly sampled K labeled examples per class from the training dataset as the support set $\mathcal{D}_{train-support}$ ($K * C$ support examples in total), and N query examples from the rest of training data as query set $\mathcal{D}_{train-query}$, output prediction values for query examples and minimized the loss on the query set to update the meta-learning model. Note that the K labeled support examples in the testing dataset $\mathcal{D}_{test-support}$ did not participate in the training stage, but just served as model input in testing tasks.

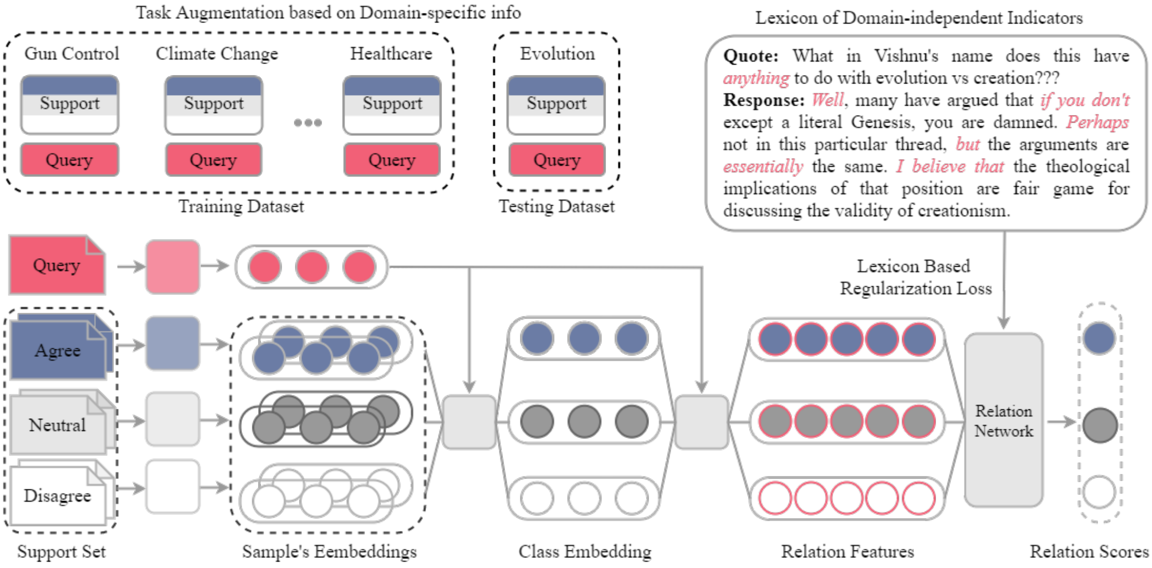
3.1.2 Attentive Class Embedding Building

Within a training/testing task, each class embedding is derived attentively from the given support examples via learned attention weights on them: first obtained the Q-R pair embedding for each support & query sample, then mapped support examples through two-layer neural networks learned separately for each class, lastly calculated the attention weights to derive attentive class embedding.

The initial embedding for the support $s_k^i, i \in \{1, \ldots, C\}, k \in \{1, \ldots, K\}$ and query examples e_q are obtained on the basis of pre-trained BERT model (Devlin et al., 2018): concatenating the hidden state vectors at the [CLS] token of quote and response sentence together as the pair embedding.

Then, we mapped support examples through a two-layer neural network learned separately for each class:

$$\hat{s}_k^i = W_2^i(W_1^i s_k^i + b_1^i) + b_2^i \tag{1}$$



Quote: What in Vishnu's name does this have *anything* to do with evolution vs creation???
Response: *Well*, many have argued that *if you don't* except a literal Genesis, you are damned. *Perhaps* not in this particular thread, *but* the arguments are *essentially* the same. *I believe that* the theological implications of that position are fair game for discussing the validity of creationism.

Figure 1: Illustration of the meta-learning approach for 3-way 2-shot problem with one query instance

At last, support examples were aggregated into class embedding via learned attentions $\{a_k^i\}_{k=1}^K$ over $\{\hat{s}_k^i\}_{k=1}^K$, in which attention weights are calculated wrt both support $\{\hat{s}_k^i\}$ and query e_q:

$$a_k^i = softmax(m^T tanh(W_3 \hat{s}_k^i + W_4 e_q)) \quad (2)$$

$$c_i = \sum_{k=1}^{K} a_k^i * \hat{s}_k^i \quad (3)$$

where $c_i, i \in \{1, \dots, C\}$ is i-th class embedding.

Different from the naive mean average in the original Prototype Network (Snell et al., 2017), our method derived class embedding from support set in an attentive way, and also took the query instance into consideration when calculating the attention weight over support examples.

3.1.3 Relation Network

With the classes embedding and query embedding at hand, the final step is to compare the query instance with each class embedding via a learned two-layer relation network, output the relation scores for each class, and choose the class with maximum relation score as the prediction result.

For each class, relation feature is designed as the concatenation of class embedding c_i, query embedding e_q, and the element-wise subtraction, element-wise multiplication, L2 norm, dot product of them:

$$f_{iq} = [c_i; e_q; c_i - e_q; c_i \odot e_q; ||c_i - e_q||; c_i \cdot e_q] \quad (4)$$

Then relation features were fed into a two-layer relation network to learn the relation scores between the i-th class and query e_q as output:

$$r_{iq} = sigmoid(W_6(W_5 f_{iq} + b_5) + b_6) \quad (5)$$

Output relation score r_{iq} is a scalar between 0 and 1 to measure the similarity between query instance and each class, and the ground truth $y_q \in \{0, 1\}$ meaning matched class has similarity 1 & mismatched class has similarity 0. The objective function we used is the mean square error (MSE) loss on the query set:

$$L_{MSE} = \sum_{i=1}^{C} \sum_{q=1}^{N} (r_{iq} - I(y_q == i))^2 \quad (6)$$

3.2 Lexicon Based Regularization Loss

To further strengthen model's ability of quickly generalizing to a new domain, we manually created a lexicon of domain-independent (dis)agreement indicators to incorporate domain-invariant features from various seen domain. Moreover, we designed a *lexicon based regularization loss* to make the meta-learner focus more on selected domain-independent indicators.

When creating the domain-independent lexicon, we followed the similar scenario in prior work (Misra and Walker, 2013), which proved domain-independent words/phrases in cue words, agreement words, denial words, and hedge words categories are all crucial to cross-domain (dis)agreement identification. We manually inspected 695 disagreements and 141 agreements

Category	Examples
Cue words (48)	so, oh, well, just, and, because, though, as well, if, then, thus, unless, seems, also, you, uh
Agreement (145)	yes, correct, agree, accept, support, true, like, good, exactly, ok, right, clear, sure, thanks, believe, of course, make sense
Denial (278)	no, not, never, nothing, however, but, doesn't, don't, isn't, yet, none, hate, false, wrong, doubt, disagree, how can, I don't think
Hedge (25)	maybe, probably, would, could, rather, although, really, actually, wondering, possibly, essentially, anyway, somewhat, I suppose

Table 1: Examples of selected domain-invariant features

from the ten domains in IAC dev set, and selected the words/phrases belonging to discourse markers associated with stating a personal opinion (cue words), agreement markers expressing support (agreement words), denial markers showing rejection/negation (denial words), and hedges that deliberately vague/soften a claim (hedge words), which are important for human to identify agree/disagree. Besides, in order to provide better generalization, we generalized the selected phrases, e.g., *I don't think* would also result in *I don't see* being added into the lexicon (Misra and Walker, 2013). Table 1 listed examples of our selected domain-independent words/phrases.

To make the meta-learner focus more on the domain-independent features, we designed a *regularization loss* to maximize the model's attention on selected domain-invariant words. For an instance consisting of n words, the model's attention on the l^{th} word is designed as L2 norm of the gradient of model output (relation scores) wrt l^{th} word's embedding. Thus, model's attention on the words in a query instance e_q is:

$$\overrightarrow{g_q} = (||\frac{\partial r_{tq}}{\partial w_{1q}}||, ||\frac{\partial r_{tq}}{\partial w_{2q}}||, \ldots, ||\frac{\partial r_{tq}}{\partial w_{nq}}||) \tag{7}$$

where $(w_{1q}, w_{2q}, \ldots, w_{nq}) \in e_q$ and $y_q = t$. Similarly, the attention on the words in a support example s_k^t is:

$$\overrightarrow{g_{s_k^t}} = (||\frac{\partial r_{tq}}{\partial w_{1s_k^t}}||, ||\frac{\partial r_{tq}}{\partial w_{2s_k^t}}||, \ldots, ||\frac{\partial r_{tq}}{\partial w_{ns_k^t}}||) \tag{8}$$

where $(w_{1s_k^t}, w_{2s_k^t}, \ldots, w_{ns_k^t}) \in s_k^t$. Bigger gradient value means more influence on the model output, and thus means more model attention. Then, we used an indicator $I(w_1, w_2, \ldots, w_n)$ to show whether the word belongs to our selected domain-independent words set, which is a vector consisting of value 0 or 1. Finally, our *regularization loss* is designed as the dot product of gradient vector (model attention) and indicator vector:

$$\begin{aligned} Lreg = -\sum_{q=1}^{N} \Bigg\{ &\overrightarrow{g_q} \cdot I(w_{1q}, w_{2q}, \ldots, w_{nq}) \\ &+ \sum_{k=1}^{K} \overrightarrow{g_{s_k^t}} \cdot I(w_{1s_k^t}, w_{2s_k^t}, \ldots, w_{ns_k^t}) \Bigg\} \end{aligned} \tag{9}$$

where $y_q = t$. Note that we added the regularization loss on both query and support examples in a training task. The total objective loss will be:

$$L_{total} = L_{MSE} + \lambda * Lreg \tag{10}$$

where λ is a hyper-parameter.

3.3 Meta-training Task Augmentation

In our previous episodic training process, we treat the entire training dataset as tasks, meaning support set and query set are sampled from the entire training dataset for each single training task, which is also the common approach used by previous papers (Murty et al., 2021). This brings us two major problems: one is meta-learning actually needs a well-defined task distribution from which a large number of diverse tasks can be sampled to train the meta-learner, another one is the entire training dataset consisting of various domains data is also heterogeneous. Thus, sampling support & query from the entire training dataset not only limited the diversity of meta-training tasks, but also resulted in support & query examples are heterogeneous with each other, making the meta-learner harder to foster the ability of quickly adapting to new domains.

For these reasons, we proposed to augment the task distribution by decomposing the entire training dataset into several sub-datasets based on domain-specific information and sampling support & query from the same sub-dataset to form training tasks. To be detailed, for the dataset having ground-truth domain labels, we grouped all the pairs by true domain labels to form distinct training domains as sub-datasets. But for the dataset without domain labels, we made use of sentences in discussion titles which mainly contains domain-specific feature to cluster the dataset into several clusters. K-means algorithm (MacQueen, 1967) is applied on the discussion title's [CLS] token embedding from pre-trained BERT, and the number of clusters is selected by elbow method (Joshi and Nalwade, 2013). In this way, training tasks are created by sampling support & query from the same sub-datasets, and ideally the same training domain.

4 Experiments

4.1 Datasets

Internet Argument Corpus (IAC) (Walker et al., 2012) annotated Q-R pair from the website 4forums.com with (dis)agreement scores from -5 to 5, with -5 as strongly disagree and 5 as strongly agree. We transformed the average score into (dis)agreement label as previous paper did (Wang and Cardie, 2016)(Misra and Walker, 2013): [-5,-1] as *Disagreement*, (-1,1) as *Neutral*, [1,5] as *Agreement*. Also, pairs in IAC have human-annotated domain labels in a total of ten domains. Train/dev/test are split in the ratio of 7:1:2 within each domain. To evaluate the meta-learner's generalization ability among different domains in the same dataset, we did experiments of both 2-way and 3-way classification within IAC and selected Evolution, Gun Control, Gay Marriage as testing domain. While testing on Evolution, for example, the training dataset $\mathcal{D}_{train}$ is all the other domains in IAC exclude Evolution, and the support set $\mathcal{D}_{test-support}$ is sampled from Evolution train set. The statistics of the test set from these three domains are listed in Table 2. When augmenting meta-training tasks, we divided $\mathcal{D}_{train}$ by golden domain label.

Agreement by Create Debaters (ABCD) dataset (Rosenthal and McKeown, 2015) collected Q-R pair from another website createdebate.com. The (dis)agreement label is derived in this way: if the side labels of Response and Quote are the same, the relation is *Agreement*, if different, the relation is *Disagreement*, and when the author is the same for both posts or Response is directly replying to the discussion title, the relation is *Neutral*. To evaluate the model's generalization ability among different datasets, we did experiment using ABCD as training dataset $\mathcal{D}_{train}$ and also chose Evolution, Gun Control, Gay Marriage in IAC as testing domain $\mathcal{D}_{test}$. Since the relation between two different users' posts can only be *Agreement* or *Disagreement* in ABCD, we can only conduct 2-way classification generalizing from ABCD to IAC. Also, when augmenting meta-training tasks, we clustered the discussion titles in ABCD with K-means and selected five as the number of clusters by elbow method. Table 3 summarized statistics of the two datasets we used in this paper.

Domain	Agree	Neutral	Disagree
Evolution	91	202	522
Gun Control	49	110	234
Gay Marriage	24	46	102

Table 2: Statistics of test sets in IAC

Dataset	Thread	Pairs	Agree	Neutral	Disagree
ABCD	10468	128343	28111	60128	40104
IAC	1806	9980	1113	2712	6155

Table 3: Statistics of ABCD and IAC datasets

4.2 Experimental Setting

Implementation Details: To evaluate proposed meta-learning approaches, we tested our models on a new domain within the same dataset as well as on a new dataset. The number of support examples per class K is set to 5, and the query set size N is set to 15 in each meta-training task. The λ in equation (10) is set to 1. Learning rate is set to 2e-5. For 2-way and 3-way classification within IAC, we trained the model from epoch 1 to 10 and selected the best one. For 2-way classification generalizing from ABCD to IAC, the number of epoch is 2.

Evaluation The testing task in the new domain consists of a support set $\mathcal{D}_{test-support}$ and a real test set $\mathcal{D}_{test-query}$ to evaluate the prediction results on. Here, we used F1 score for each class and macro Precision/Recall/F1 score as evaluation metrics. To control for variations across different support sets, we sampled 50 random support sets for each testing task, and report the average results on these support sets.

Baseline Conventional transfer learning which trained a supervised model on the dataset with richer labeling resource and then fine tuned on the few provided labeled examples in a new domain is commonly used previously. To train a supervised model on the training dataset $\mathcal{D}_{train}$, we also used the same instance embedding as in the meta-learning models, then add a classification layer on top of it and activated by a softmax layer to output the probability for each class. The loss in supervised model is the classical cross-entropy classification loss. Then we fine tuned it on the support set $\mathcal{D}_{test-support}$ in the new testing domain, and also reported the average results on 50 randomly sampled support sets.

Test Domain	Evolution					Gun Control					Gay Marriage				
Model	A	D	Macro			A	D	Macro			A	D	Macro		
	F1		P	R	F1	F1		P	R	F1	F1		P	R	F1
Supervised	71.4	81.9	80.8	72.6	76.7	75.2	78.8	77.4	76.7	77.0	87.1	89.4	88.9	87.6	88.2
Fine Tune	77.9	77.6	78.5	77.0	77.7	76.0	77.4	77.3	76.2	76.7	86.1	86.7	86.7	86.2	86.4
Meta	76.6	83.6	82.3	77.9	80.1	80.9	80.0	80.5	80.4	80.4	91.7	93.1	93.1	91.7	92.4
Meta + reg	79.0	85.3	**84.6**	79.7	82.1	81.4	83.0	82.3	82.1	82.2	93.6	94.5	94.5	93.7	94.1
Meta + aug	80.0	84.4	83.1	81.3	82.2	84.8	82.1	84.0	82.9	83.4	93.7	94.6	94.6	93.7	94.2
Meta + aug + reg	**81.0**	**85.3**	84.3	**82.1**	**83.2**	**85.3**	**83.2**	**84.7**	**83.9**	**84.3**	**95.8**	**96.3**	**96.3**	**95.9**	**96.1**

Table 4: Results of 2-way classification when training on other domains exclude test domain within IAC (A: Agreement, D: Disagreement, P; Precision, R: Recall, F1: F1 score)

Test Domain	Evolution						Gun Control						Gay Marriage					
Model	A	N	D	Macro			A	N	D	Macro			A	N	D	Macro		
		F1		P	R	F1		F1		P	R	F1		F1		P	R	F1
Supervised	50.4	49.1	55.6	54.9	48.5	51.7	52.3	36.4	52.0	48.2	45.6	46.9	66.3	50.8	57.8	62.7	53.9	58.3
Fine Tune	54.4	47.5	56.8	55.2	50.5	52.9	56.2	38.2	51.8	49.6	47.8	48.7	66.0	51.8	58.6	61.8	55.8	58.8
Meta	61.4	51.1	56.6	60.0	52.6	56.4	56.1	47.1	53.7	55.7	48.9	52.3	69.1	54.4	63.8	64.5	60.4	62.4
Meta + reg	63.5	50.2	61.8	60.1	56.9	58.5	59.2	**49.8**	53.0	57.4	50.5	54.0	70.6	**59.1**	65.6	66.4	63.8	65.1
Meta + aug	64.0	52.5	62.0	60.6	58.4	59.5	58.9	46.2	**57.7**	56.7	51.9	54.3	77.3	44.8	**67.3**	63.1	63.1	63.1
Meta + aug + reg	**64.3**	**53.9**	**64.7**	**62.6**	**59.4**	**61.0**	**60.8**	47.7	57.2	**58.4**	**52.0**	**55.2**	**78.4**	56.7	64.1	**67.3**	**65.4**	**66.4**

Table 5: Results of 3-way classification when training on other domains exclude test domain within IAC (A: Agreement, N: Neutral, D: Disagreement, P; Precision, R: Recall, F1: F1 score)

4.3 Results

Experimental results are summarized in Table 4-6. For the baseline methods, we reported the supervised model's performance before and after fine tuning in the first (Supervised) and second (Fine Tune) row. For our meta-learning approaches, we reported the performance of basic meta-learning model (Meta), only adding the lexicon based regularization loss (Meta + reg), only with task augmentation (Meta + aug), and task augmentation together with regularization loss (Meta + aug + reg).

Table 4-5 shows the experiments generalizing among domains within the same dataset, we can observe that fine tuning on the few labeled examples can bring us F1 score improvement in most cases, but sometime it will lower the performance. As analyzed in (Zhang et al., 2017), it is due to the overfitting problem for transfer learning without large amount of training data, so that the model simply memorize the labeled samples and fail to learn generalizable features for new domains. In comparison, under both 2-way and 3-way classification, the basic meta-learning model can outperform fine tuning baseline by 2.4% to 6% in macro F1 score across all the three testing domains consistently. The meta-learning with lexicon based regularization and task augmentation can further enhance the domain adaptation ability, resulting in 2.9% to 4.6% increase in F1 score compared to the basic meta-learner, in accompany with improvement in both Precision and Recall.

Table 6 evaluates the models' ability to generalize across different datasets. Since the Q-R pairs in ABCD and IAC are collected from different website sources, the distribution of data points in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are distinct with each other, making this experiment setting more difficult. Similar to the experiments within the same dataset, fine tuning also has the problem of overfitting and received a poor performance under Evolution and Gun Control domain. The basic meta-learning model shows better performance than fine tuning baseline across all the testing domains, with 1.8% to 2.6% increase in macro F1 score. After training with lexicon based regularization & task augmentation, the performance was further increased by 2.6% to 6.5% in macro F1 score, with improvement in both Precision and Recall. The above results indicate that the basic meta-learner can achieve better adaptation performance, while incorporating lexicon based regularization loss and task augmentation together can further boost meta-learner's domain generalization ability.

4.4 Ablation Study

Effectiveness of lexicon based regularization loss Comparing Meta + reg with Meta, the results show making meta-learner focus more on domain-independent features through lexicon based regularization can generate better performance across all

Test Domain	Evolution					Gun Control					Gay Marriage				
Model	A	D	Macro			A	D	Macro			A	D	Macro		
	F1		P	R	F1	F1		P	R	F1	F1		P	R	F1
Supervised	77.1	82.3	80.6	78.8	79.7	76.8	76.8	76.8	76.8	76.8	84.1	83.5	83.9	83.8	83.8
Fine Tune	77.0	81.5	80.1	78.4	79.2	76.6	76.5	76.9	76.2	76.5	84.8	86.1	85.8	85.1	85.4
Meta	79.7	82.2	81.2	80.8	81.0	80.8	77.4	79.5	78.6	79.1	86.4	88.1	87.6	87.1	87.3
Meta + reg	81.3	84.6	83.5	82.5	83.0	80.7	**81.9**	81.4	81.3	81.3	89.2	88.9	89.1	89.0	89.0
Meta + aug	80.7	84.2	83.0	81.9	82.4	82.1	78.8	81.1	79.9	80.5	90.0	90.8	90.6	90.3	90.4
Meta + aug + reg	**82.8**	**85.9**	**84.9**	**83.7**	**84.3**	**82.3**	81.1	**81.8**	**81.6**	**81.7**	**93.3**	**94.3**	**94.3**	**93.3**	**93.8**

Table 6: Results of 2-way classification when training on ABCD and testing on domains in IAC (A: Agreement, D: Disagreement, P; Precision, R: Recall, F1: F1 score)

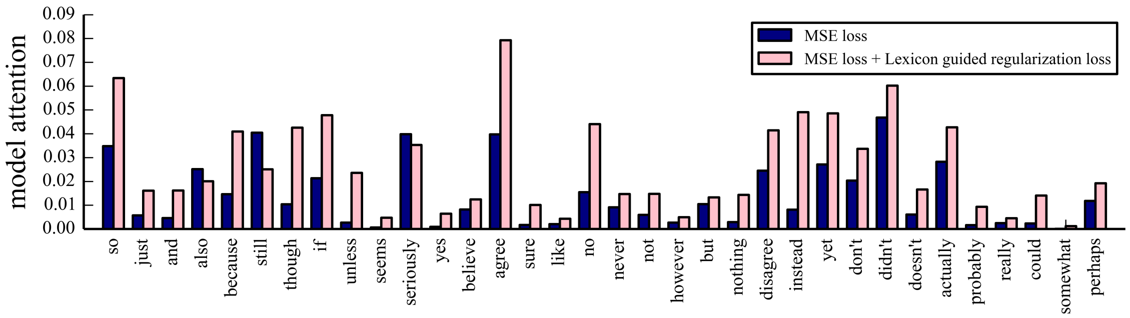


Figure 2: Change of model attention on selected domain-independent words after lexicon based regularization loss

the testing domains, in both Precision and Recall, resulting in 0.9% to 3.4% macro F1 increase.

Effectiveness of task augmentation Comparing Meta + aug with Meta, we can see task augmentation will also bring us improvement, ranging from 0.7% to 3.1% F1 score raise, in terms of both Precision and Recall.

In addition, combing the regularization loss and task augmentation together will generate the most improvement wrt the basic meta-learner, with 2.6% to 6.5% increase in macro F1 score. Meta + aug + reg can outperform the fine tuning baseline by a total number of 5.1% to 9.7% F1 score.

4.5 Analysis

We found that after training with additional regularization loss on seen domains, the meta-learner will indeed focus more on our selected domain-independent words in unseen domains during testing. Take 2-way classification within IAC dataset as an example, the average of model attentions on our selected domain-independent words was increased by 0.6%, 0.7%, 1.2% when testing on Evolution, Gun Control, Gay Marriage domain respectively. This observation is also valid on other experiments: the lexicon based regularization loss will increase the model attention on domain-independent words from 0.4% to 1.2%, thus can indeed make meta-learner focus more on them.

We also analyzed the model attention on each single domain-independent word before and after adding the regularization loss. Figure 2 shows some examples in the experiment of 2-way classification within IAC while testing on Gay Marriage domain. We can observe that the additional regularization loss can help some (dis)agreement related words like "so", "because", "though", "agree", "no", "disagree", "yet", "don’t", "didn’t", "actually", "perhaps" seize more model attention on them and thus generate better results. There are also few words receiving less attention, like "still", "also", "seriously", we can interpret this as these words are actually weaker indicators for meta-learner.

5 Conclusion

In this paper, we developed a metric-based meta-learning model for few-shot (dis)agreement identification problem. To enhance the meta-learner’s domain generalization, we firstly manually created a lexicon of domain-independent (dis)agreement indicators and designed a lexicon based regularization loss, secondly augmented task distribution by decomposing the entire dataset into different training domains based on domain-specific info to form diverse types of meta-training tasks. In the future work, we plan to refine the manually created lexicon and replace some weak words/phrases with stronger indicators.

References

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 151–159, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Ruiying Geng, Binhua Li, Yongbin Li, Yuxiao Ye, Ping Jian, and Jian Sun. 2019. Few-shot text classification with induction network. *CoRR*, abs/1902.10482.

Xiang Jiang, Mohammad Havaei, Gabriel Chartrand, Hassan Chouaib, Thomas Vincent, Andrew Jesson, Nicolas Chapados, and Stan Matwin. 2018. Attentive task-agnostic meta-learning for few-shot text classification.

Kalpana D. Joshi and Prakash S. Nalwade. 2013. Modified k-means for better initial cluster centres.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition.

Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2020. Exploiting the matching information in the support set for few shot event classification. *CoRR*, abs/2002.05295.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Amita Misra and Marilyn Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France. Association for Computational Linguistics.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA. PMLR.

Shikhar Murty, Tatsunori Hashimoto, and Christopher D Manning. 2021. Dreca: A general task augmentation strategy for few-shot natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1113–1125.

Janarthanan Rajendran, Alexander Irpan, and Eric Jang. 2020. Meta-learning requires meta-augmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 5705–5715. Curran Associates, Inc.

Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Lu Wang and Claire Cardie. 2016. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *arXiv preprint arXiv:1606.05706*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization.