Biomedical Relation Classification by Single and Multiple source Domain Adaptation

Sinchani Chakraborty [1] **Sudeshna Sarkar** [2] **Pawan Goyal** [2] **Mahanandeeshwar Gattu** [3]
[1,2]Department of Computer Science and Engineering, IIT Kharagpur
[3]Excelra Knowledge Solutions Pvt Ltd, Hyderabad, India
[1] sinchanichakraborty@gmail.com
[2]{sudeshna, pawang}@cse.iitkgp.ac.in

Abstract

Relation classification is crucial for inferring semantic relatedness between entities in a piece of text. These systems can be trained given labelled data. However, relation classification is very domain-specific and it takes a lot of effort to label data for a new domain. In this paper, we explore domain adaptation techniques for this task. While past works have focused on single source domain adaptation for bio-medical relation classification, we classify relations in an unlabeled target domain by transferring useful knowledge from one or more related source domains. Our experiments with the model have shown to improve state-of-the-art F1 score on 3 benchmark biomedical corpora for single domain and on 2 out of 3 for multi-domain scenarios. When used with contextualized embeddings, there is further boost in performance outperforming neural-network based domain adaptation baselines for both the cases.

1 Introduction

In the biomedical domain, a relation can exist between various entity types like protein-protein, drug-drug, chemical-protein etc. Detecting relationships is a fundamental sub-task for automatic Information Extraction to overcome efforts of manual inspection, especially for growing biomedical articles. However, existing supervised systems are highly data-driven. This poses a challenge since manual labelling is a costly and time-consuming process and there is a dearth of labelled data in the biomedical domain covering all tasks and for new datasets. A system trained on a specific dataset[1] may perform poorly on another, for the same task (Mou et al., 2016), due to dataset variance which can arise owing to sample selection bias (Rios et al., 2018).

[1]Note: We use the terms dataset and domain interchangeably.

Domain Adaptation aims at adapting a model trained on a source domain to another target domain that may differ in their underlying data distributions. Past work on domain adaptation for bio-medical relation classification has focused on single-source adaptation (Rios et al., 2018). However, multiple sources from related domains can prove to be beneficial for classification in a low-resource scenario.

In this paper, we perform domain adaptation for biomedical binary relation classification at the sentence-level. For single-source single target (SSST) we transfer between different datasets of protein-protein interaction, along with drug-drug interaction. We also explore multi-source single target (MSST) adaptation to incorporate more richness in the knowledge transferred by using additional smaller corpora for protein-protein relation and multiple labels for chemical-protein relation respectively. Given an unlabeled target domain, we transfer common useful features from related labelled source domains using adversarial training (Goodfellow et al., 2014). It helps to overcome the sampling bias and learn common indistinguishable features, promoting generalization, using min-max optimization. We adopt the Multinomial Adversarial Network integrated with the Shared-Private model (Chen and Cardie, 2018) which was originally proposed for the task of Multi-Domain Text Classification. It can handle multiple source domains at a time which is in contrast to traditional binomial adversarial networks. The Shared-Private model (Bousmalis et al., 2016) consists of a split representation where the private space learns specific features related to a particular domain while a shared space learns features common to all the domains. Such representation promotes non-contamination of the two spaces preserving their uniqueness. The contributions of our approach are as follows:

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 75–80
Hong Kong, November 3, 2019.
https://doi.org/10.18653/v1/D19-62

1) We show that using a shared-private model along with adversarial training improves SSST adaptation compared to neural network baselines. When multiple source corpora from similar domains are used it leads to further performance enhancement. Moreover, use of contextualized sentential embeddings leads to better performance than exisitng baseline methods for both MSST and SSST.

2) We explore the generalizability of our framework using two prominent neural architectures: CNN (Nguyen and Grishman, 2015) and Bi-LSTM (Kavuluru et al., 2017), where we find the former to be more robust across our experiments.

2 Methodology

For every labeled sources and a single unlabeled target we have set of NER tagged sentences, each of which is represented as: X = $\{e_1, e_2, w_1 ... w_n\}$ where e_1 and e_2 are two tagged entities and w_j is the j^{th} word in the sentence . A labelled source instance is accompanied by the relation label (True or False). In this section we discuss the input representation followed by model description.

2.1 Input Representation

We form word and position embeddings for every word in an NER tagged sentence. We use the PubMed-and-PMC-w2v[2] to generate word embeddings. The size being $(|V| \cdot d_w)$, where d_w is the word embedding dimension which is 200 and $|V|$ is the vocabulary size. The position embedding vector for j^{th} word in a sentence relative to two tagged entities e_1 and e_2 is represented as a tuple: $(p_{e1(j)}, p_{e2(j)})$ where, $p_{e1(j)}$ and $p_{e2(j)} \in \mathbb{R}^e$.

2.2 Model

Fig 1 shows the adaptation of MAN framework whose various components are discussed below.

Shared & Domain feature extractor (F_s, F_{d_i}) The input representation is fed to both F_{d_i} and F_s for labeled source domains whereas for unlabeled target instances it is fed only to F_s. For SSST the model is trained on a single labeled source domain and tested on a unlabeled target domain. For MSST we do not combine the sources as a single corpus since that leads to a number of false negatives. We make two different assumptions to consider multiple sources: 1) Following Nguyen et al., (2014) we consider multiple labels from single corpus as different sources, 2) We use additional smaller corpora from a similar domain as multi-source. The *Shared Feature space* (F_s) learns domain agnostic representations and *Private Feature space* (F_{d_i}) learns domain specific features for every i^{th} labeled domain. We apply two different architectures for both F_s and F_{d_i} to analyze the changes in performance of the approach for the task : Convolutional neural network (Nguyen and Grishman, 2015) (MAN CNN) and Bi-LSTM (Kavuluru et al., 2017) (MAN Bi-LSTM). We have performed detailed experiments on each of these in Section 5.

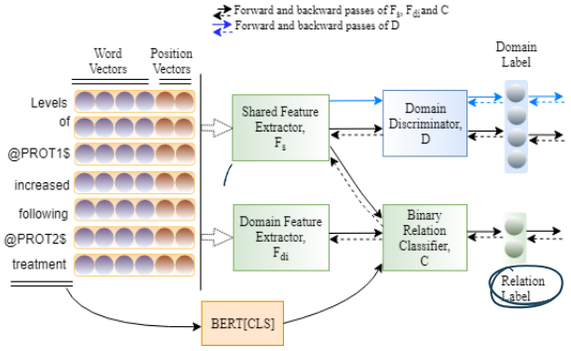

Figure 1: MAN for Domain Adaptation of Binary Relation Classification. The figure shows the training flow given a sentence from a labeled source domain. D is trained separately than rest of the network

Domain discriminator, D is a fully-connected layer with softmax that predicts multiple domain probabilities using *Multinomial Adversarial Network*. The output from F_s is fed to D which is adversarially trained separately from the entire network using L2 loss described as follows:

$$L_D(\hat{d}, d) = \sum_{i=1}^{N} (\hat{d}_i - 1_{\{d=i\}})^2$$

where, d is the index assigned for a domain and $\hat{d}$ is the prediction. It is generalized as $\sum_{i=1}^{N} \hat{d}_i = 1$ and $\forall_i : \hat{d}_i \geq 0$. F_s tries to fool D so that it can not correctly guess the domain from where a sample instance is coming from. Thus F_s learns indistinguishable features in the process.

Relation Classifier C is a fully-connected layer with a softmax, used to predict the class probabilities. We use Bio-BERT (Lee et al., 2019) embeddings for every sentence as features (Geeticka Chauhan, 2019) BERT[CLS] that have shown to improve performance in many downstream tasks. This is concatenated with the fixed size sentence representation from F_s and F_{d_i}, together

[2]http://evexdb.org/pmresources/vec-space-models/

Datasets	Entity Pair	# of Sent	# of Positive	# of Negative
AiMed	PPI	1995	1000	4834
BioInfer		1100	2534	7132
LLL		77	164	166
HPRD50		145	163	270
IEPA		486	355	482

Table 1: Protein Protein Interaction Dataset statistics.

Datasets	Entity Pair	# of Train	# of Valid	# of Test
DDI	Drug-Drug	27779	-	5713
CPR: 3	Chem-Prot	768	550	665
CPR: 4		2254	1094	1661
CPR: 5		173	116	195
CPR: 6		235	199	293
CPR: 9		727	457	644

Table 2: Drug-Drug Interaction and Chemical-Protein Dataset statistics.

they serve as input to C. For unlabeled target, during test no domain specific features are generated from F_{d_i} and that part is set to zero vector. For binary classification we adopt Negative Log Likelihood Loss for C described below:

$$L_c(\hat{y}, y) = -logP(\hat{y} = y)$$

where, y is the true relation label and $\hat{y}$ is the softmax label. The objective of F_{d_i} is same as that of C and it relies only on labeled data. On the other hand the objective of the Shared Feature Extractor F_s is represented as follows:

$$Loss\, of\, F_s = Classifier\, loss + \lambda\, Domain\, loss$$

It consists of two loss components: improve performance of C and enhance learning of invariant features across all domains. A hyper parameter λ is used to balance both of them.

3 Datasets

The dataset statistics is summarized in Table 1 and Table 2. A 10-fold cross validation was performed for the Protein-Protein Interaction dataset. For given set of entities E in a sentence, it is split into $\begin{pmatrix} E \\ 2 \end{pmatrix}$ instances. All positive instances of datasets with more than two relation types are merged and assigned True labels while negative instances are assigned False labels. Unlabeled data is formed by removing labels from development and test datasets.

4 Experiments

Pre-processing: We anonymize the named entities in the sentence by replacing them with predefined tags like @PROT1$, @DRUG$ (Bhasuran and Natarajan, 2018).

4.1 Single source single target (SSST)

A thorough experiment is conducted using all possible combinations of the three benchmark data-sets AiMed (Bunescu et al., 2005), BioInfer (Pyysalo et al., 2006), DDI (Herrero-Zazo et al., 2013) whose results are discussed in Table 3

4.2 Multi-source single target (MSST)

The experiments with two different assumptions to consider multiple sources are as follows:

Multiple smaller corpora from similar domain: For Protein Protein Interaction there are three smaller standard corpora in literature, namely, LLL (Nedellec, 2005), IEPA (Ding et al., 2001), HPRD50 (Fundel et al., 2007). All three were considered as additional sources to transfer knowledge. AiMed (AM) and BioInfer (BI) were alternately selected as the unlabeled target in 2 different experiments while the remaining 4 denoted as 4P are considered as source corpus.

Multiple labels from single corpora: For ChemProt corpora we consider various labels as different sources following Nguyen et al., (2014) The five positive labels of ChemProt are: CPR: 3, CPR: 4, CPR: 5, CPR: 6, CPR: 9 which stand for upregulator, downregulator, agonist, antagonist and substrate, respectively. We predict the classification performance for unlabeled targets CPR:6 and CPR:9 taking multi-source labeled input denoted as 3C from three sources- CPR: 3, CPR: 4, CPR: 5 as positive instances and remaining as negative.

4.3 Baselines

We compare our approach with different baselines which are mentioned as follows:

- **BioBERT (Rios et al., 2018):** For SSST we train it on one dataset and test on another. For MSST we combine the multiple sources as a single source and test on labeled target.

- **CNN+DANN (Lisheng Fu, 2017) :** A variant of adversarial training which is gradient reversal (RevGrad) is used with CNN (Nguyen and Grishman, 2015).

Method	BioInfer → AiMed	AiMed → BioInfer	BioInfer → DDI	DDI → BioInfer	AiMed → DDI	DDI → AiMed
CNN	45.22	36.72	39.75	22.13	15.83	27.93
Bi-LSTM	46.88	29.59	40.87	17.21	18.58	25.80
BioBERT*	76.48	69.23	67.89	57.84	51.22	54.83
CNN + DANN*	45.98	42.01	41.58	34.37	28.66	28.90
Bi-LSTM + RevGrad	46.41	40.11	39.41	37.20	27.72	35.29
Adv-CNN	48.79	54.13	44.19	48.53	45.96	44.71
Adv - Bi-LSTM	48.51	56.54	44.47	44.90	46.21	43.44
MAN CNN **	50.23	55.04	47.63	49.51	46.97	42.36
MAN Bi-LSTM **	49.19	58.69	46.77	46.28	47.84	41.53
MAN CNN + BERT[CLS] **	**53.08**	57.89	49.33	50.79	47.01	46.38
MAN Bi-LSTM + BERT[CLS] **	52.74	**61.01**	48.03	45.12	**50.19**	44.01

Table 3: F1 scores for SSST experiment on test set of target (RHS of →) . **: Our model. *: Our implementation. Bold text: Best domain adaptation model for a dataset.

- **Adv Bi-LSTM + Adv CNN (Rios et al., 2018):** Conducts two-step training: pre-training with source followed by adversarial training with target. For MSST experiment we compare our method with Adv CNN and Adv Bi-LSTM by combining multiple sources.

5 Results and Discussions

In Table 3 we see that BioInfer generalizes well to AiMed and DDI corpora using vanilla LSTM or CNN architecture. However, with MAN and contextual embeddings, we do not see prominent gains as much as the other datasets. This can be due to the class imbalance in data (positive to negative instance ratio 1:5.9) (Hsu et al., 2015; Rios et al., 2018). For AiMed and BioInfer, we find that the knowledge transfer among themselves gives the best performance thus strengthening the fact that datasets from the same domain can contribute to performance enhancement justifying the performance gains in MSST experiments. Our model outperforms other baselines just with the use of adversarial training which might be attributed to joint learning better representation from shared and private feature extractors. The use of contextual BERT[CLS] tokens leads to increase in performance scores since they encode important relations between words in a sentence (Vig, 2019; Hewitt and Manning, 2019).

In Table 4, BioBERT is seen to perform well for ChemProt. We hypothesize that this may be due to the same underlying dataset being used during train and test. Though we use different labels as multi-source, that may not contribute to generating enough variance in sources since they were from the same dataset. For AiMed and BioInfer, however, three different smaller corpora were used, where the proposed method outperforms BioBERT. When compared across all the six SSST experiments, the Bi-LSTM based model lacks in performance may be due to absence of any attention mechanism which would have helped in selecting more relevant context (Chen and Cardie, 2018). We observe that adversarial training along with contextualized BERT sentence embeddings leads to performance gains across all datasets.

Method	3C → CPR:9	3C → CPR:6	4P → AM	4P → BI
BioBERT*	**69.27**	**73.50**	43.01	52.98
Adv-CNN*	58.23	56.69	45.30	51.79
Adv-BiLSTM*	56.30	57.13	42.01	52.67
MAN CNN**	59.69	58.30	52.33	57.21
MAN Bi-LSTM**	57.01	59.71	53.64	59.37
MAN CNN + BERT -[CLS]**	64.23	65.41	56.75	**64.83**
MAN Bi-LSTM + BERT -[CLS]**	62.07	64.09	**57.09**	63.92

Table 4: F1 scores for MSST experiment on test set of target (RHS of →). **: Our model. *: Our implementation trained with unified labeled multi-source. Bold text: Best model for a dataset..

6 Conclusions

Our proposed model significantly outperformed the existing neural network based domain adaptation baselines for SSST. Among the two MSST experiments, we showed that the system gains when multiple source corpora are used. We also experiment with two architectures out of which CNN is seen to perform marginally better compared to Bi-LSTM. Our analysis on Section 5 further explains the effect of sources, adversarial training and use of contextualized BERT sentential embeddings.

Acknowledgments

This work has been supported by the project Effective Drug Repurposing through literature and patent mining, data integration and development of systems pharmacology platform sponsored by MHRD, India and Excelra Knowledge Solutions, Hyderabad. Besides, the authors would like to thank the anonymous reviewers for their valuable comments and feedback.

References

Balu Bhasuran and Jeyakumar Natarajan. 2018. Automatic extraction of gene-disease associations from literature using joint ensemble learning. In *PloS one*.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *NIPS*.

Razvan C. Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine*, 33 2:139–55.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of NAACL-HLT 2018*, page 12261240.

Jing Ding, Daniel Berleant, Dan Nettleton, and Eve Syrkin Wurtele. 2001. Mining medline: Abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 326–37.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relex - relation extraction using dependency parse trees. *Bioinformatics*, 23 3:365–71.

Peter Szolovits Geeticka Chauhan, Matthew B. A. McDermott. 2019. Reflex: Flexible framework for relation extraction in multiple domains. In *Proceedings of the BioNLP 2019 workshop*, page 3047.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of biomedical informatics*, 46 5:914–20.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL-HLT*.

Tzu-Ming Harry Hsu, Wei-Yu Chen, Cheng-An Hou, Yao-Hung Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. 2015. Unsupervised domain adaptation with imbalanced cross-domain data. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4121–4129.

Ramakanth Kavuluru, Anthony Rios, and Tung Tran. 2017. Extracting drug-drug interactions with word and character-level recurrent neural networks. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 5–12.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *ArXiv*, abs/1901.08746.

Bonan Min Ralph Grishman Lisheng Fu, Thien Huu Nguyen. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, page 425429.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yuning Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? *ArXiv*, abs/1603.06111.

Claire Nedellec. 2005. Learning language in logic - genic interaction extraction challenge.

Minh Luan Nguyen, Ivor Wai-Hung Tsang, Kian Ming Adam Chai, and Hai Leong Chieu. 2014. Robust domain adaptation for relation extraction via clustering consistency. In *ACL*.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *VS@HLT-NAACL*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2006. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50 – 50.

Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Generalizing biomedical relation classification with neural adversarial domain adaptation. volume 34, pages 2973–2981.

Jesse Vig. 2019. Visualizing attention in transformer-based language representation models. *ArXiv*, abs/1904.02679.