Knowledge Distillation for Discourse Relation Analysis

Congcong Jiang[1], Tieyun Qian[1], Bing Liu[2]*

[1]School of Computer Science, Wuhan University, China

[2]Department of Computer Science, University of Illinois at Chicago, USA

Introduction

Discourse relation recognition (DRR) aims to identify the discourse relations that hold between two text spans. It consists of explicit and implicit discourse relation recognition (termed as EDRR and IDRR), whose difference depends on whether the connectives like 'as' exist or not in the data.

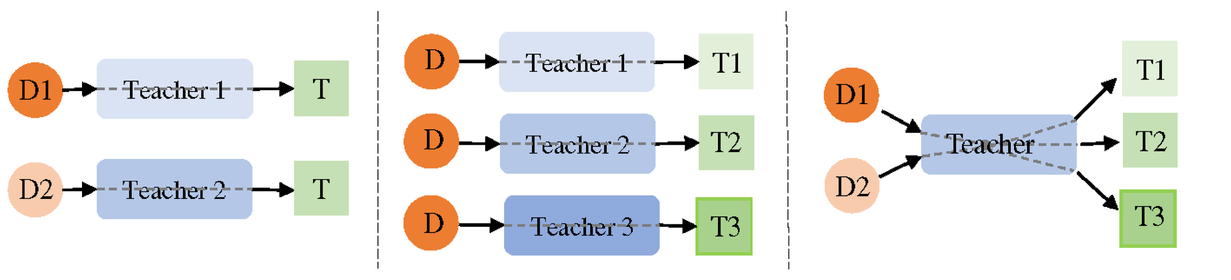
Challenge:

The problems of linguistic dissimilarity and different class distributions [1] make it hard to get optimal performance by directly training EDRR and IDRR with Multi-task Learning Framework (MTL).

Motivation:

- **Our goal is to retain the benefit of MTL in acquiring the common knowledge across data or tasks, and to exploit KD's power to transfer knowledge from a multi-data multi-task teacher to a single-data single-task student.**
- **The Knowledge Distillation framework of Multi-teacher Multi-task (MTMT) or Multi-teacher Multi-data (MTMD) is not always necessary when the difference between tasks/data is not significant, e.g., geometry and algebra. The implicit and explicit data in our study also belong to this case.**

Our Contributions



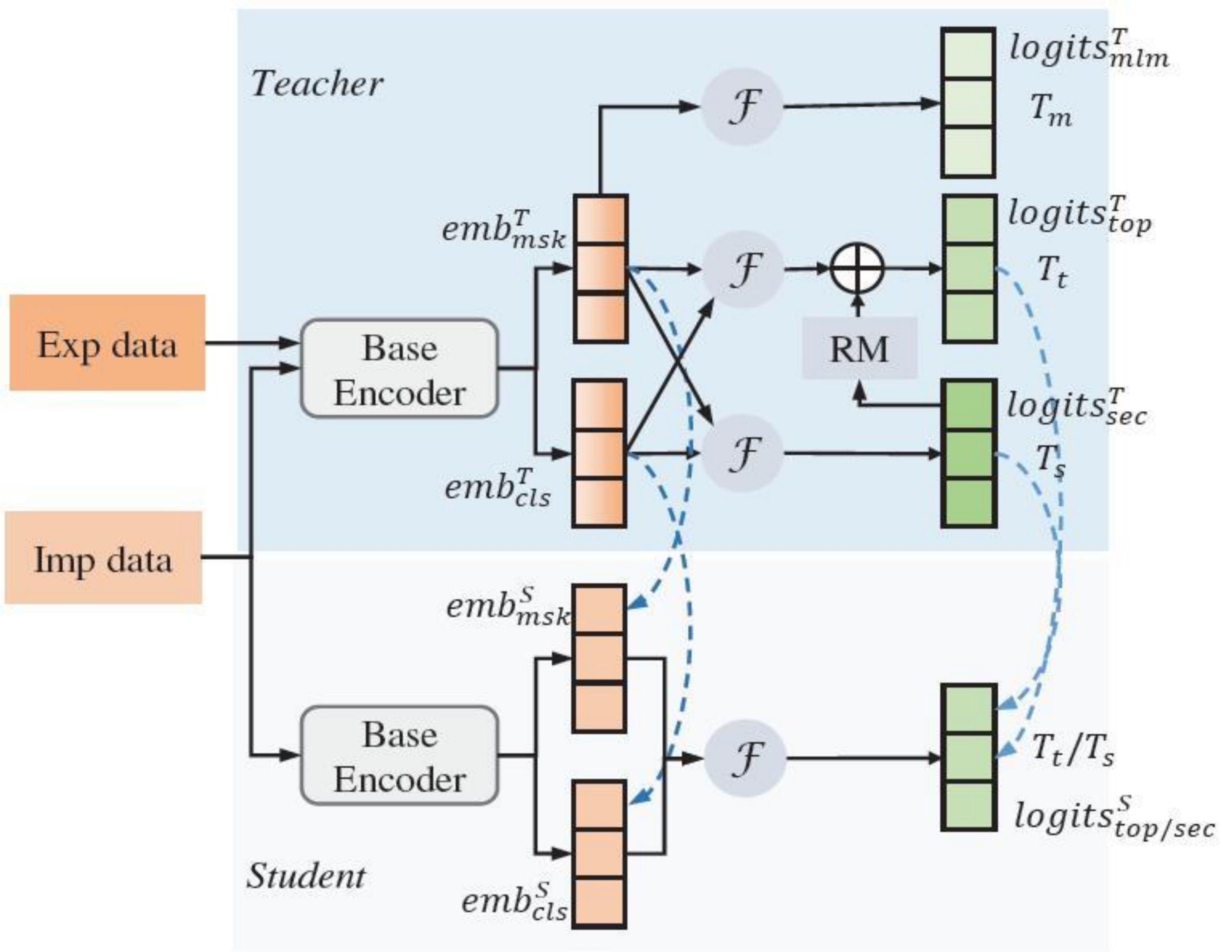
(a) multi-teacher multi-data (b) multi-teacher multi-task (c) one-teacher multi-data multi-task

OTMT: One-Teacher Multi-task Multi-data

- **We develop a novel one-teacher multi-data multi-task (OTMT for short) knowledge distillation framework for the IDRR task as shown in (c).**
- **From the data perspective, one general teacher trained on different data can enhance the model's adaptivity to data.**
- **From the task perspective, one general teacher trained for different tasks can enforce the model to learn the connections, including the shared parameter space and the public features among tasks.**
- **From the complexity perspective, one general teacher shares the data and the encoder structure in the same parameter space. As a result, it has the benefit of a small model size and also avoids the complicated ensemble procedure of multiple teacher models.**

OTMT Network

An Overview of OTMT



- **Teacher Network: We use explicit and implicit data, and perform the top level and second level relation classification and an additional auxiliary masked language modeling (MLM) task to train the teacher network.**

  <u>The Base Encoder</u>: We adopt several Pre-trained Language Models (PLMs) including BERT, RoBERTa, and XLNet as the base encoder for both the teacher and student networks.

  <u>The Relation Matrix (RM)</u>: If a relation belongs to the second level class, it must belong to the corresponding top level class too, and the entry in RM is 1 otherwise 0.

- **Student Network: The student model takes implicit data as the only input, and trains one network for the top and the second level classification task separately. The student model adopts the base encoder with the same structure and same size as that in teacher network.**
- **Knowledge from Teacher to Student: To effectively transfer knowledge from the general multi-data and multi-task teacher to the single-data and single-task student networks, we propose to exploit two types of information learned by the teacher model including the soft labels and the feature vectors.**

  <u>Training Loss for Student Network</u>: In order to train each student network, we need to optimize the prediction and knowledge distillation targets at the same time.

Training Procedure

- **We save the teacher network that performs the best on the validation set of implicit relation data.**
- **We then generate the corresponding soft labels and feature vectors for implicit samples, and use them together with the ground truth labels to guide the student network training.**

Experimental Evaluation

Experimental Results

Model	Top		Second (Acc)		
	Acc	F1	Lin	Ji	P&K
M1 [11](Bb)*	66.12	57.42	52.13	52.43	52.72
M2 [9](Bb)	66.01	57.17	52.12	52.32	52.34
M3 [6](Rb)*	67.14	57.84	52.38	55.39	55.15
M4 [3](Bb)	65.52	56.27	51.94	51.89	51.88
M4 [3](Bl)*	68.30	60.61	54.36	56.23	55.12
M4 [3](Xb)*	66.35	59.33	54.33	54.62	54.36
M4 [3](Xl)*	69.52	63.58	57.44	59.51	58.21
OTMT (Bb)	**66.94**	**59.19†**	**54.15†**	**53.65†**	**53.67†**
OTMT (Bl)	**70.02‡**	**61.35†**	**56.03‡**	**57.55†**	**56.99†**
OTMT (Rb)	**70.54‡**	**62.27‡**	**56.87‡**	**58.02‡**	**57.17†**
OTMT (Xb)	**68.89‡**	**60.78†**	**56.37‡**	**56.65‡**	**56.95‡**
OTMT (Xl)	**72.34‡**	**64.46†**	**61.62‡**	**61.06†**	**61.56‡**

	Top		Second (Acc)			Complexity	
	Acc	F1	Lin	Ji	P&K	Time	Space
OTMT	**66.94**	**59.19**	**54.15**	**53.65**	**53.67**	$1.18h$	$222M$
MTL	61.66‡	51.11‡	50.65‡	48.41‡	50.05‡	$\mathbf{1.11}h$	$\mathbf{110}M$
MTMD	66.12	58.00	52.64†	52.38	53.20	$1.87h$	$332M$
MTMT	65.43	56.76‡	52.17†	52.97	53.18	$2.64h$	$394M$
w/o stu.	61.66	51.11	50.65	48.41	50.05	-	-
w/o tea.	65.49	55.45	51.07	52.61	52.17	-	-
w/o s.l.	66.38	57.50	52.40	52.96	53.67	-	-
w/o f.v.	66.37	57.71	52.01	52.78	52.61	-	-

- **Dataset: PDTB 2.0 [2].**
- **Lin/Ji/P&K: 3 ways to split the dataset.**
- **(Bb) = (BERT-base), (Bl) = (BERT-large).**
- **(Xb) = (XLNet-base), (Xl) = (XLNet-large).**
- **(Rb)=(RoBERTa-base), (Rl) = (RoBERTa-large).**
- ***h* = hour, *M* = 1×10^6.**

Conclusion

- **We propose a novel one-teacher multi-data multi-task KD framework. Better than multi-task learning, our model leverages the KD's ability of transferring knowledge from a general teacher model to a specific student model.**
- **Different from multi-teacher KD, our model shares the common knowledge across multiple data and multiple tasks using one-teacher network with the low computational cost.**
- **Extensive experimental results on the popular PDTB dataset prove that our model significantly outperforms both the state-of-the-art baselines and the variants with the multi-task learning or multi-teacher KD architecture.**

References

- **[1] Lan et al. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition. ACL 2013.**
- **[2] Prasad et al. The Penn Discourse TreeBank 2.0. LREC 2008.**