A Label Dependence-aware Sequence Generation Model for Multi-level Implicit Discourse Relation Recognition

Changxing Wu[1], Liuwen Cao[1], Yubin Ge[2], Yang Liu[3], Min Zhang[4], Jinsong Su[5]*

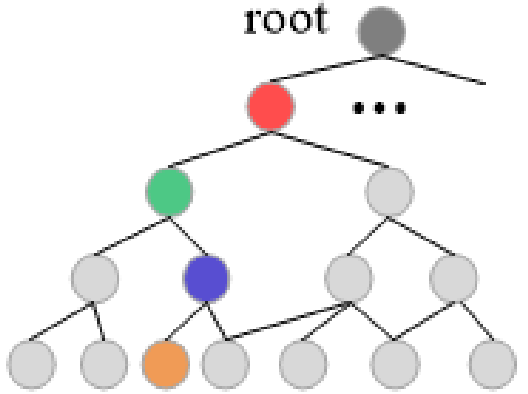
1.East China Jiaotong University 2. Illinois University

3.Tsinghua University 4.Soochow University 5.Xiamen University

Background

[我今天不出去打篮球了,]arg_1
[今天外面下雨.]arg_2

- The top-level label: *Contingency*
- The second-level label: *Cause*
- The third-level label: *Reason*
- The inserted connective: *because*



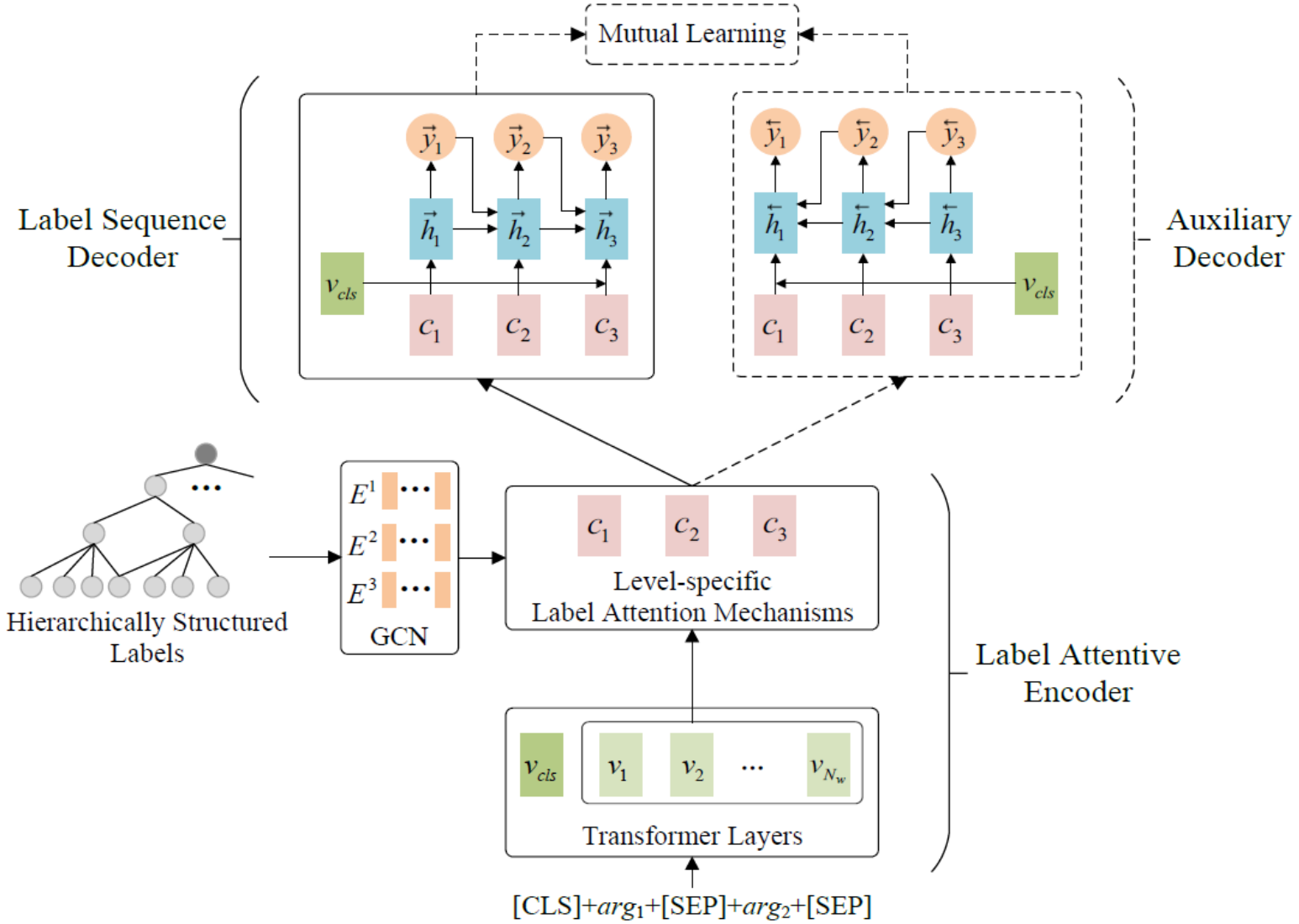
Most existing methods train multiple models to predict multi-level labels independently, while ignoring the dependence between hierarchically structured labels.

Main Contribution

- We consider multi-level IDRR as a label sequence generation task. To our knowledge, our work is the first attempt to deal with this task in the way of sequence generation.

- We propose a label dependence-aware sequence generation model. Both its encoder and decoder, combined with the mutual learning enhanced training method, can fully leverage the label dependence for multi-level IDRR.

LDSGM Model


Mutual Learning
Label Sequence Decoder
Auxiliary Decoder
Hierarchically Structured Labels
GCN
Level-specific Label Attention Mechanisms
Label Attentive Encoder
Transformer Layers
[CLS]+arg1+[SEP]+arg2+[SEP]

Label Attentive Encoder

- Local and global representations

$$v_{cls}, v_1, ..., v_i, ..., v_{N_w} = \text{Transformers}(\text{arg}_1, \text{arg}_2)$$

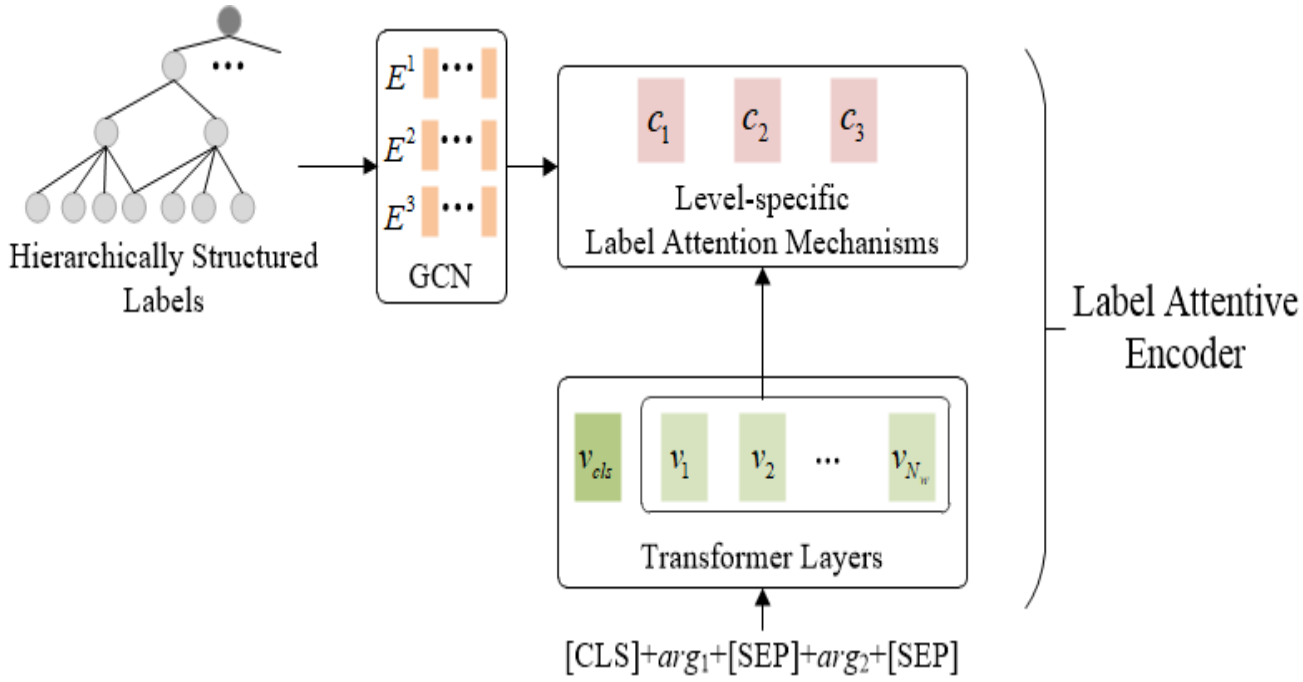
- Use GCN to obtain better label embeddings

$$e_j^l = \sigma(\sum_{k=1}^{N_C} A_{jk} W^l e_k^{l-1} + b^l)$$

- Label attention mechanisms to extract level-specific contexts

$$c_m = V\alpha$$

$$\alpha = \text{softmax}(\text{max-pooling}(V^T W^m E^m))$$



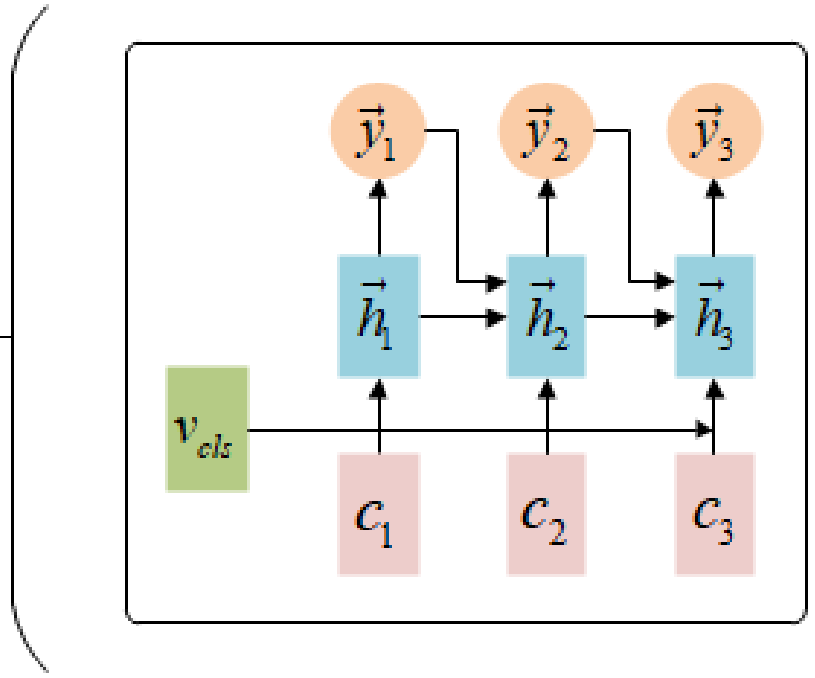
Label Sequence Decoder

Our decoder is an RNN-based one that sequentially generates the predicted labels in a top-down manner, that is, the top-level label, the second-level label, and so on. By doing so, the easily-predicted higher-level labels can be used to guide the label prediction at the current level.

$$\vec{y}_m = \text{softmax}(W_o \vec{h}_m + b_o)$$

$$\vec{h}_m = \text{GRU}(\vec{h}_{m-1}, [v_{cls}; c_m; g(\vec{y}_{m-1})])$$

Label Sequence Decoder

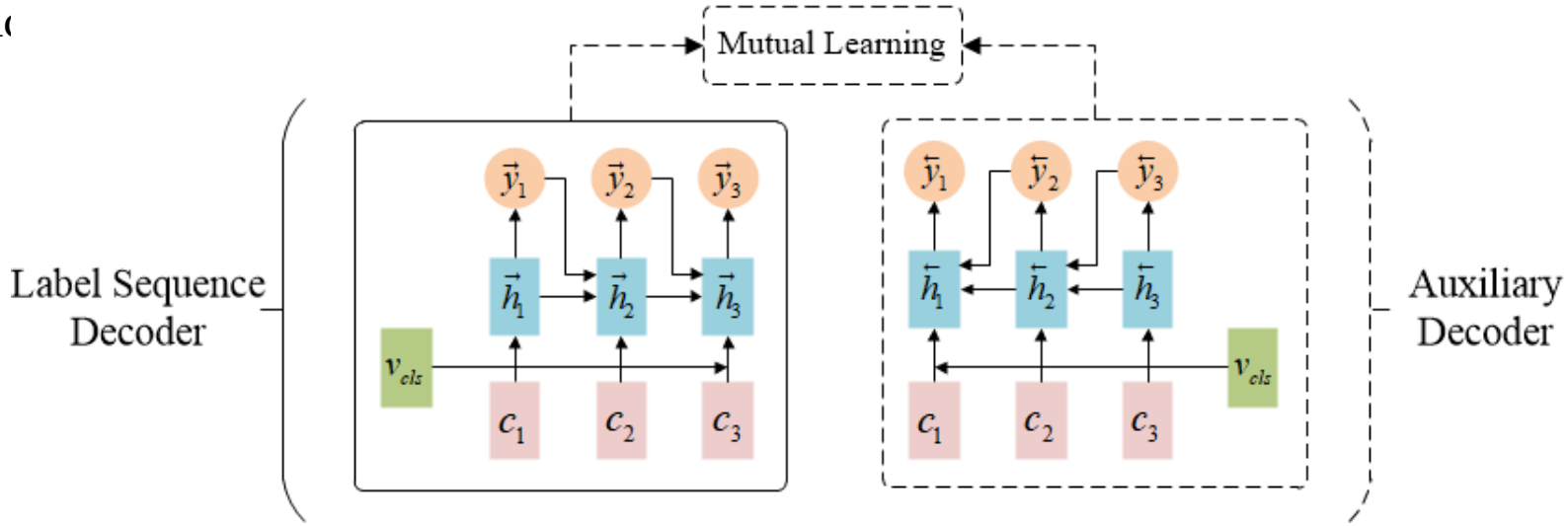


Mutual Learning Enhanced Training

our decoder generates labels in a top-down manner, which can only exploit the dependence from the predicted higher-level labels, leaving that from lower levels unexploited. The basic intuition behind our auxiliary decoder is that annotators usually insert a suitable connective to help the higher-level label annotations of a given instan

$$L(D;\theta_e,\theta_d)=\sum_{(x,y)\in D}\sum_{m=1}^{M}\{-E_{y_m}[\log \overset{r}{y}_m]+\lambda*\mathrm{KL}(\overset{s}{y}_m \,\mathrm{P}\, \overset{r}{y}_m)\}$$

$$L(D;\theta_e;\theta_{ad})=\sum_{(x,y)\in D}\sum_{m=1}^{M}\{-E_{y_m}[\log \overset{s}{y}_m]+\lambda*\mathrm{KL}(\overset{r}{y}_m \,\mathrm{P}\, \overset{s}{y}_m)\}$$



Result

Model	Embeddings	Top-level		Second-level		Connective	
		F_1	Acc	F_1	Acc	F_1	Acc
IDRR-C&E (Dai and Huang 2019)	GloVe	50.49	58.32	32.13	46.03	-	-
IDRR-Con (Shi and Demberg 2019a)	word2vec	46.40	61.42	-	47.83	-	-
KANN (Guo et al. 2020)	GloVe	47.90	57.25	-	-	-	-
IDRR-C&E (Dai and Huang 2019)	ELMo	52.89	59.66	33.41	48.23	-	-
MTL-MLoss (Nguyen et al. 2019)	ELMo	53.00	-	-	49.95	-	-
BERT-FT (Kishimoto, Murawaki, and Kurohashi 2020)	BERT	58.48	65.26	-	54.32	-	-
HierMTN-CRF (Wu et al. 2020)	BERT	55.72	65.26	33.91	52.34	10.37	30.00
BMGF-RoBERTa (Liu et al. 2020)	RoBERTa	63.39	69.06	-	58.13	-	-
MTL-MLoss-RoBERTa	RoBERTa	61.89	68.42	38.10	57.72	7.75	29.57
HierMTN-CRF-RoBERTa	RoBERTa	62.02	70.05	38.28	58.61	10.45	31.30
OurEncoder+OurDecoder	RoBERTa	62.93	70.66	39.71	59.59	10.67	31.54
LDSGM	RoBERTa	**63.73**	**71.18**	**40.49**	**60.33**	**10.68**	**32.20**

Experimental results on PDTB.

Analysis

Unlike most previous studies predicting labels at different levels independently, our LDSGM model converts multi-level IDRR into a label sequence generation task, which intuitively is able to alleviate the inconsistency among label predictions at different levels

Model	Top-Sec	Top-Sec-Con
HierMTN-CRF	46.29	19.15
BMGF-RoBERTa	47.06	21.37
OurEncoder+OurDecoder	57.86	25.31
LDSGM	**58.61**	**26.85**

Comparison with recent models on the consistency among multi-level label predictions.

Ablation Study

LA means the level-specific label attention mechanism, PP represents previous predictions, and ML means the mutual learning enhanced training. Softmax+MultiTask denotes that we stack softmax layers on the top of our encoder to individually predict labels at different levels, and then train them based on multi-task learning.

Model	Top-level		Second-level		Connective	
	F_1	Acc	F_1	Acc	F_1	Acc
LDSGM	**63.73**	**71.18**	**40.49**	**60.33**	**10.68**	**32.20**
w/o GCN	63.47	70.77	39.32	59.54	9.81	30.94
w/o LA	62.84	70.19	39.15	59.09	10.64	31.32
w/o PP	63.68	70.78	39.27	59.62	10.08	30.86
w/o ML (aka. OurEncoder+OurDecoder)	62.93	70.66	39.71	59.59	10.67	31.54
OurEncoder+Softmax+MultiTask	62.82	70.25	38.28	58.37	9.60	30.82
OurEncoder+CRF	62.82	70.53	38.82	59.09	10.41	31.24
OurEncoder+Ensemble(OurDecoder, AuxDecoder)	63.04	70.80	39.91	59.61	8.61	29.95

Ablation study

Analysis

The performance of the labels which have few annotated instances is still poor. In the future, we will pay more attention to the minority labels

Second-level Label	BMGF-RoBERTa	LDSGM
Temp.Asynchronous	56.18	**56.47**
Temp.Synchrony	0.0	0.0
Cont.Cause	59.60	**64.36**
Cont.Pragmatic cause	0.0	0.0
Comp.Contrast	59.75	**63.52**
Comp.Concession	0.0	0.0
Expa.Conjunction	60.17	57.91
Expa.Instantiation	67.96	**72.60**
Expa.Restatement	53.83	**58.06**
Expa.Alternative	60.00	**63.46**
Expa.List	0.0	**8.98**

Label-wise F1 scores for the second-level labels.

Thanks