Graph Adaptive Semantic Transfer for Cross-domain Sentiment Classification

Kai Zhang
Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
kkzhang0808@mail.ustc.edu.cn

Qi Liu, Zhenya Huang
Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
{qiliuql,huangzhy}@ustc.edu.cn

Mingyue Cheng
Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
mycheng@mail.ustc.edu.cn

Kun Zhang
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
zhang1024kun@gmail.com

Mengdi Zhang, Wei Wu
Meituan
Beijing, China
mdzhangmd@gmail.com
wuwei19850318@gmail.com

Enhong Chen*
Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China
Hefei, China
cheneh@ustc.edu.cn

(SIGIR-2022)

Reported by Zhaoze Gao

Artificial

1. Introduction
2. Approach
3. Experiments

gesis
Leibniz-Institut für Sozialwissenschaften

Introduction

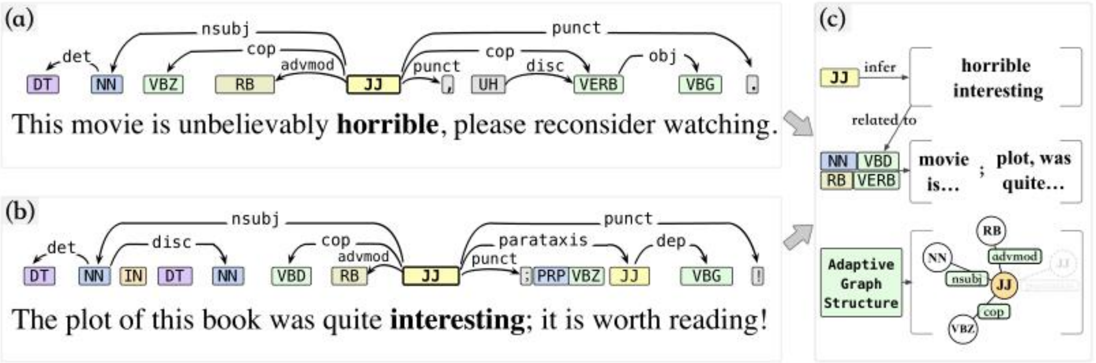


Figure 1: The transferable syntactic structures* of two examples (i.e., (a), (b)). The colorful boxes ("DT") and black lines (e.g., "det") indicate POS tags and syntactic relations, respectively. As shown in (c), the syntactic structures are similar between domains so that it is easy for human to understand the hidden knowledge behind sentences in different domains. However, those adaptive graph features are largely ignored by existing domain adaptation research.

First, sentiment words play a crucial role in CDSC, while POS tags can distinguish sentiment words (e.g., "horrible" and "interesting" in Figure 1) via the POS tag "JJ" in a natural way, i.e., the "JJ" label means the word is an adjective.

Second, the sentiment polarity of reviews is largely influenced by the sentiment word's neighbors, whether they are in-domain or across-domain.

Third,the syntactic graph structures of sentences in different domains are remarkably similar, which means that the syntactic rules are domain-invariant and can be naturally transferred across domains.

Approach

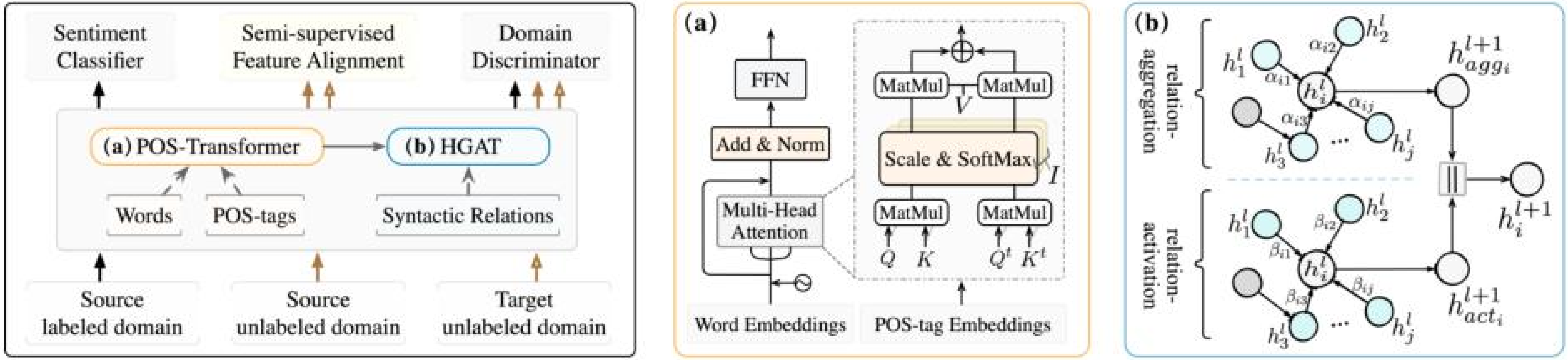
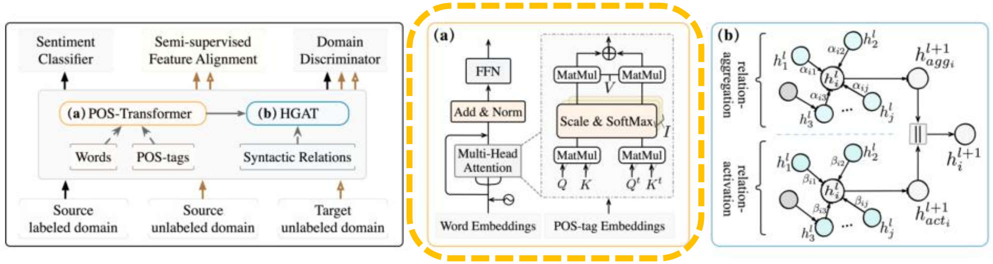
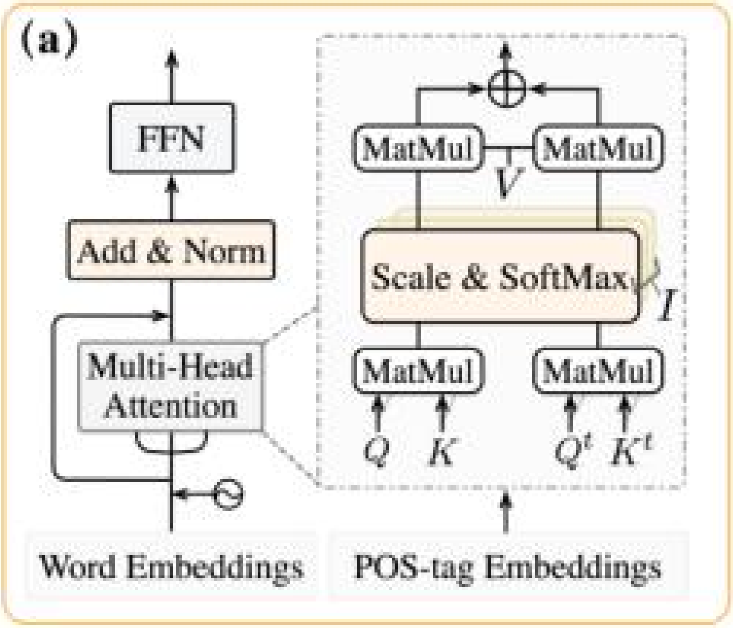


Figure 2: The architecture of GAST, which consists three parts: (a) the *POS-Transformer* that can learn sequential semantic representation by considering both the word sequences and POS tags; (b) the *HGAT* module which can exploit adaptive syntactic semantics of the sentence through the syntactic relation graph. (c) an *IDS* (i.e., Sentiment Classifier, Semi-supervised Feature Alignment and Domain Discriminator) to optimize the model and encourage it to be domain-invariant and syntax-aware.

Approach





$\{x_s^i, y_s^i\}_{i=1}^{n_{sl}} \in \mathcal{D}_s^l \quad \{x_s^i\}_{i=n_{sl}+1}^{n_s} \in \mathcal{D}_s^u)$

$\mathcal{D}_t = \{x_t^i\}_{i=1}^{n_t}$

$s = \{s_1, s_2, ..., s_n\}$

$\mathcal{G} = (\mathcal{V}, \mathcal{A}, \mathcal{R})$

$\mathcal{A}$ is adjacent matrix with $A_{ij} = 1$ if there
exists a dependency relation between word s_i and s_j, and $A_{ij} = 0$
$\mathcal{R}$ is a set of syntactic relations (e.g., *det*, *nsubj* and *cop*)

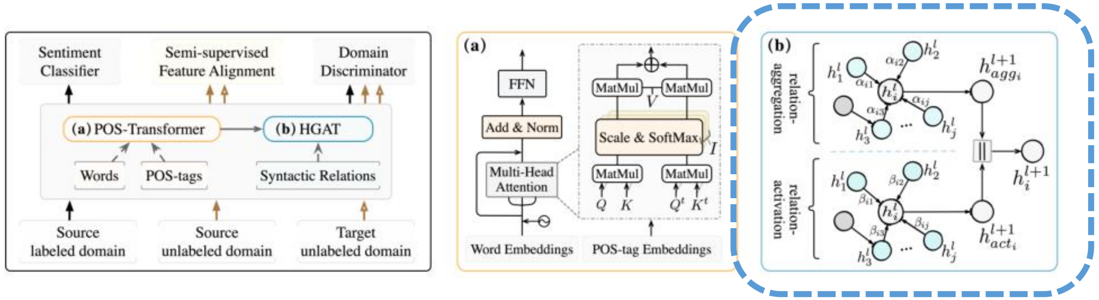
$$Z = concat(z_1, z_2, \ldots, z_I), \tag{1}$$

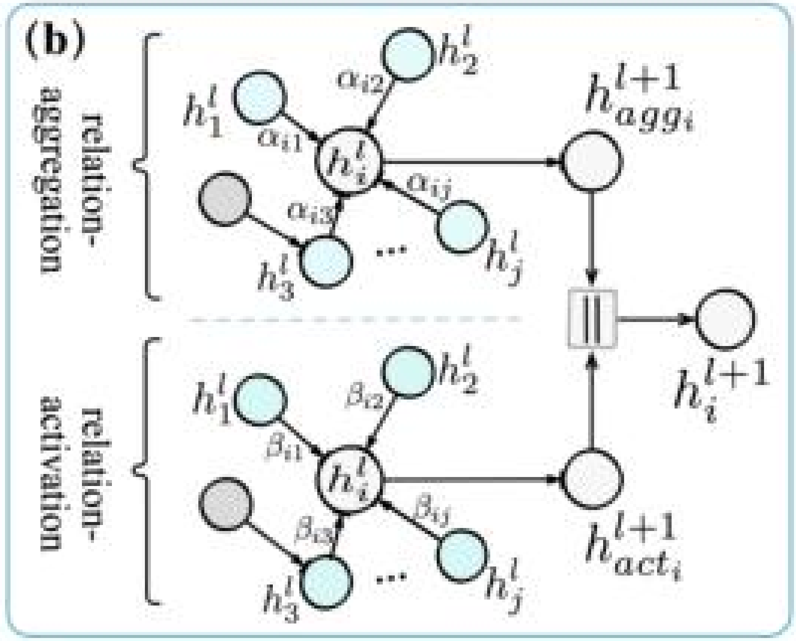
$$z_i = Att.(Q_i, K_i, V_i) + Att.(Q_i^t, K_i^t, V_i), \tag{2}$$

$$Att.(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d/I}}\right)V, \tag{3}$$

$$R = \max(0, ZW_1 + b_1)W_2 + b_2, \tag{4}$$

Approach





$$h_{agg_i}^{l+1} = \|_{k=1}^{\bar{K}} \sigma\Big(\sum_{j\in N_i} \alpha_{ij}^{lk} W_{lk} h_j^l\Big), \tag{5}$$

$$f_{ij}^{lk} = \sigma(a_{lk}^T [W_{lk}h_i^l \| W_{lk}h_j^l \| W_{lk} r_{ij}]), \tag{6}$$

$$\alpha_{ij}^{lk} = \frac{\exp\left(f_{ij}^{lk}\right)}{\sum_{j=1}^{N_i} \exp\left(f_{ij}^{lk}\right)}, \tag{7}$$

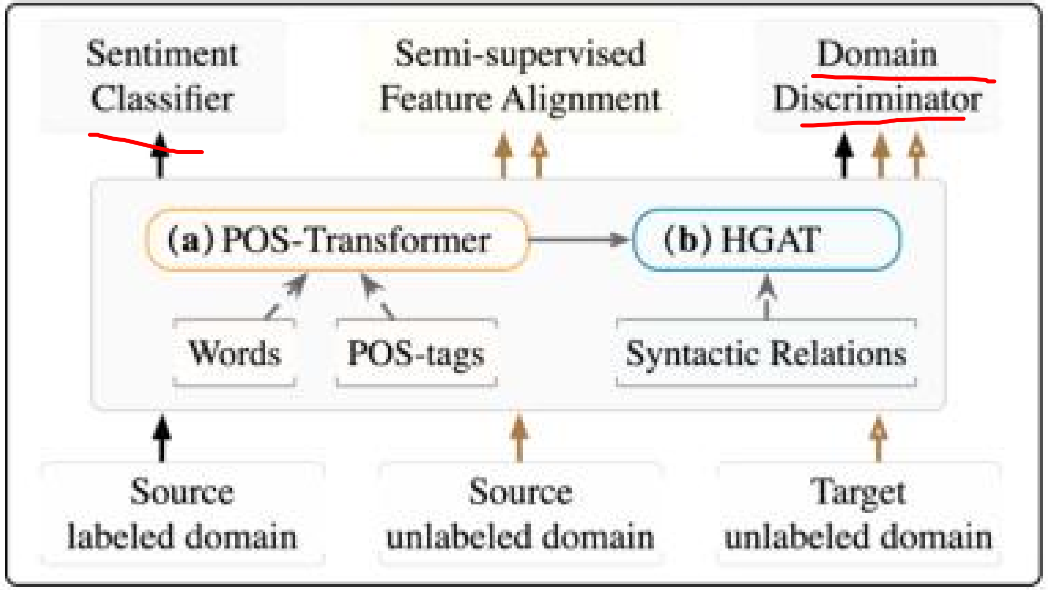
$$\beta_{ij}^{lk} = \frac{\exp\left(F_{act.}(h_i^l, h_j^l)\right)}{\sum_{j=1}^{N_i} \exp\left(F_{act.}(h_i^l, h_j^l)\right)}, \tag{8}$$

$$F_{act.} = \frac{\left(W_Q^{lk} h_i^l\right)\left(W_K^{lk} h_j^l + W_{Kr}^l r_{ij}\right)^T}{\sqrt{d/\bar{K}}}, \tag{9}$$

$$h_{act_i}^{l+1} = \|_{k=1}^{\bar{K}} \sigma\Big(\sum_{j\in N_i} \beta_{ij}^{lk} (W_V^{lk} h_j^l + W_{Vr}^l r_{ij})\Big), \tag{10}$$

$$h_i^{l+1} = h_{agg_i}^{l+1} \,\|\, h_{act_i}^{l+1}. \tag{11}$$

Approach



$$L_c = -\frac{1}{n_s^l}\sum_{i=1}^{n_s^l}(y_s^i \ln \hat{y}_s^i + (1-y_s^i)\ln(1-\hat{y}_s^i)), \qquad (12)$$

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}(y_d^i \ln \hat{y}_d^i + (1-y_d^i)\ln(1-\hat{y}_d^i)), \qquad (13)$$

$$L_a = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C} \tilde{y}^i \ln \tilde{y}^i, \qquad (14)$$

$$L = \lambda_c L_c + \lambda_d L_d + \lambda_a L_a, \qquad (15)$$

Approach

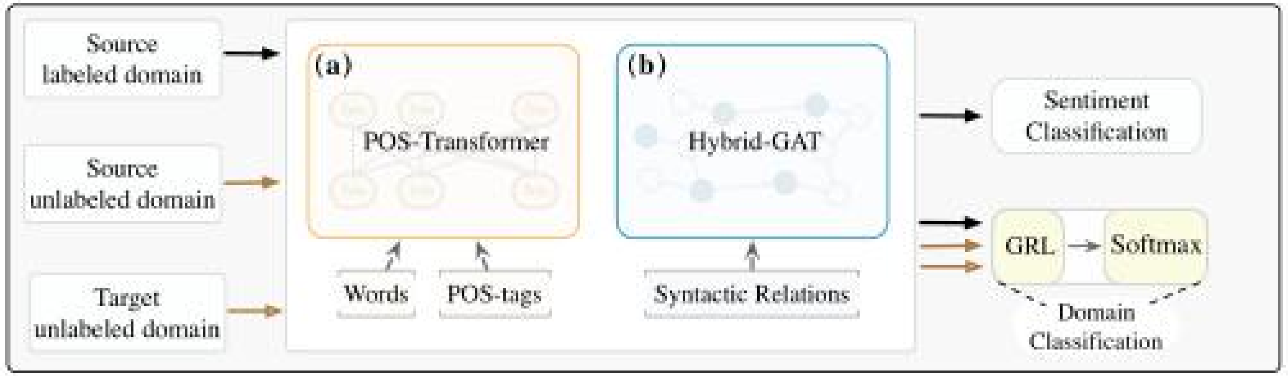


Figure 7: The framework of the ablation model G_*Non_IDS* as described in section 4.6. It mainly includes two tasks, i.e., sentiment classification and domain classification.

$$\hat{y}^d = softmax(W_d R + b_d) . \tag{16}$$

$$G(x) = x , \quad \frac{\partial G(x)}{\partial x} = -I . \tag{17}$$

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}\left(y_d^i \ln \hat{y}_d^i + \left(1 - y_d^i\right)\ln\left(1 - \hat{y}_d^i\right)\right), \tag{18}$$

Experiments

Table 1: Statistics of datasets after pre-processing.

Domains	Testing set percentage			
	#Train	#Vali.	#Test	#Unlabel
Books	1,600	400	2,000	4,000
DVD	1,600	400	2,000	4,000
Electronics	1,600	400	2,000	4,000
Kitchen	1,600	400	2,000	4,000

Experiments

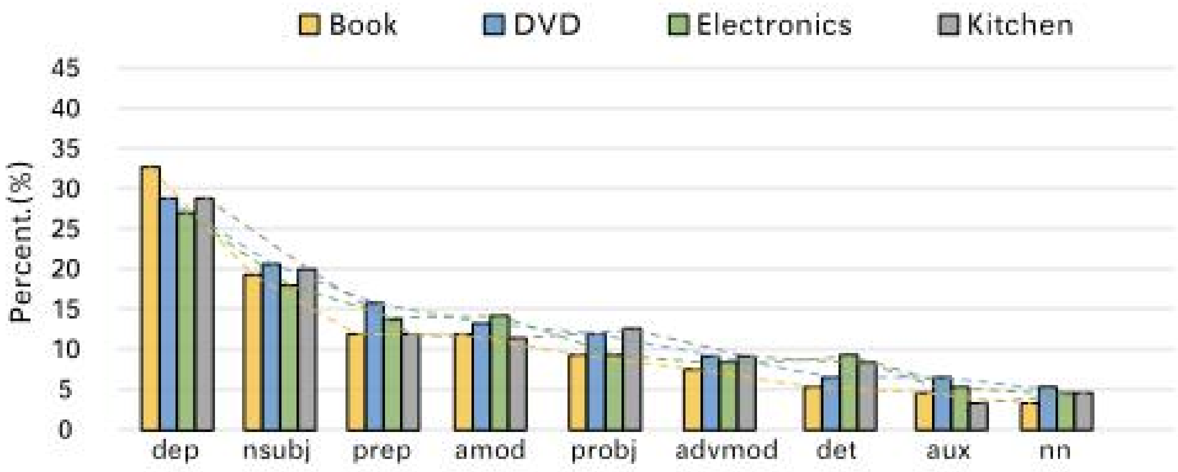


Figure 3: The percent of transferable dependency relations in different domains. We visualized the top 9 relations.

Experiments

Table 2: Sentiment classification accuracy (%) on the twelve transfer tasks.

Baselines	*DVD* (*D*)			*Book* (*B*)			*Electronics* (*E*)			*Kitchen* (*K*)		
	$D \mapsto B$	$D \mapsto E$	$D \mapsto K$	$B \mapsto D$	$B \mapsto E$	$B \mapsto K$	$E \mapsto D$	$E \mapsto B$	$E \mapsto K$	$K \mapsto D$	$K \mapsto B$	$K \mapsto E$
SCL	77.8	75.2	75.5	80.4	76.5	77.1	74.5	71.6	81.7	75.2	71.3	78.8
SFA	78.8	75.8	75.7	81.3	75.6	76.9	75.4	72.4	82.6	74.7	72.4	80.7
DANN	80.5	77.6	78.8	83.2	76.4	77.2	77.6	73.5	84.2	75.1	74.3	82.2
AMN	84.5	81.2	82.7	85.6	82.4	81.7	81.7	76.6	85.7	81.5	80.9	86.1
HATN	86.6	86.3	87.4	86.5	85.7	86.8	84.3	81.5	87.9	84.7	84.1	87.0
IATN	87.0	86.9	85.8	86.8	86.5	85.9	84.1	81.8	88.7	84.4	84.7	87.6
BERT-DAAT	90.8	89.3	90.5	89.7	89.5	90.7	90.1	88.9	93.1	88.8	87.9	91.7
LSTM	75.6	73.4	-	78.6	75.2	-	72.2	69.6	-	-	-	-
TextGCN	80.8	77.6	79.2	85.3	81.1	79.7	82.6	78.2	82.3	83.3	84.1	81.7
FastGCN	81.6	80.6	81.1	86.0	82.7	82.0	83.5	78.7	84.5	84.2	85.7	83.4
GAST	87.9	87.3	89.1	88.2	86.2	87.4	85.6	83.4	89.3	87.7	87.5	89.4
BERT-GAST	**91.1**	**90.7**	**92.1**	**90.4**	**91.2**	**91.5**	**90.7**	**89.4**	**93.5**	**89.7**	**89.2**	**92.6**
G_*Non_Pos-Tran.*	85.9	84.7	87.6	86.8	83.4	85.5	84.2	80.4	87.8	85.8	85.5	87.4
G_*Non_HGAT*	86.6	85.9	88.1	87.4	85.0	86.1	84.5	81.3	88.2	86.4	86.7	88.2
G_*Non_IDS*	87.2	86.6	87.9	87.6	85.8	86.7	85.0	82.6	88.5	85.9	86.2	87.7
G_*Non_agg*	87.5	86.7	88.9	88.0	85.9	86.9	85.2	82.6	89.0	87.3	87.2	89.1
G_*Non_act*	87.3	86.3	88.7	87.7	85.3	86.2	84.8	81.8	88.7	86.9	87.1	88.7

Experiments

	This	movie	is	unbelievably	**horrible**,	please	reconsider	watching.
(1) Transformer	0.08	0.11	0.09	0.16	0.15	0.19	0.13	0.09
(2) Pos-Transformer	0.02	0.17	0.11	0.21	0.18	0.06	0.15	0.10
(3) Hybrid GAT	0.00	0.22	0.12	0.26	0.21	0.00	0.19	0.00

(a) Case Example 1: The attention values related to the word “horrible”.

	The	plot	of	this	book	was	quite	**interesting**;	it	is	worth	reading!
(1) Transformer	0.05	0.08	0.03	0.06	0.14	0.07	0.13	0.18	0.05	0.03	0.14	0.04
(2) Pos-Transformer	0.03	0.09	0.02	0.02	0.12	0.07	0.17	0.21	0.02	0.05	0.12	0.08
(3) Hybrid GAT	0.00	0.16	0.00	0.00	0.00	0.21	0.28	0.24	0.00	0.00	0.11	0.00

(b) Case Example 2: The attention values related to the word “interesting”.

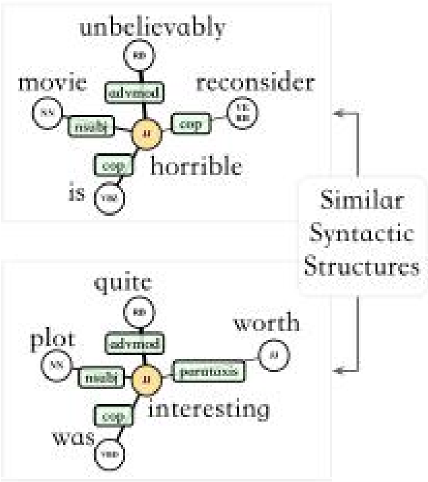


Figure 4: Attention score visualization of the different words. The attention values from vanilla attention (i.e., $Att.(Q, K, V)$ in formula 2), POS-attention (i.e., $Att.(Q^t, K^t, V)$ in formula 2) and HGAT (i.e., β in formula 8) are associated with the row (1), row (2), and row (3) respectively in both examples. Note that, some values are infinitely close to 0. That makes sense because HGAT makes the attention value more concentrated on the syntactic-related words.

Experiments

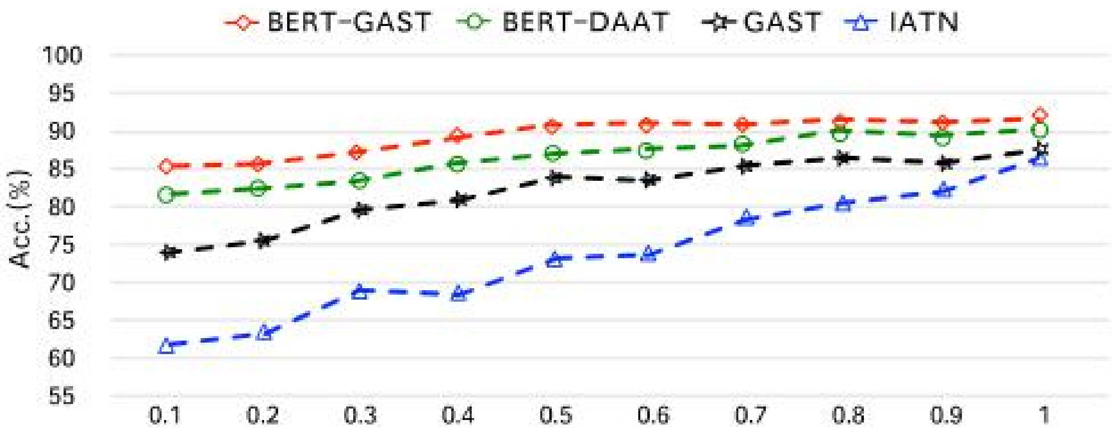


Figure 5: The influence of sample number. We explore the impact of sample number with different ratio (i.e., abscissa) of source domain. For the limited space, we only show the results of the task "$B \mapsto D$".

Experiments

Table 3: The performance (%) of different syntactic graphs constructed by different parsers on $D\mapsto*$ tasks.

Syntax Parser	$D\mapsto B$	$D\mapsto E$	$D\mapsto K$
(1) Without Graph	86.6	85.9	88.1
(2) Stanford Graph	**87.1**	**86.6**	**88.6**
+compare with (1)	**(+0.5)**	**(+0.7)**	**(+0.5)**
(3) Biaffine Graph	**87.9**	**87.3**	**89.1**
+compare with (1)	**(+1.3)**	**(+1.4)**	**(+1.0)**
+compare with (2)	**(+0.8)**	**(+0.7)**	**(+0.5)**

Experiments

Table 4: The Influence of model depth (i.e., attention heads) on $D\mapsto *$ tasks. The metric is accuracy (%).

Models	$D\mapsto B$	$D\mapsto E$	$D\mapsto K$
HGAT w 1 head	86.9	86.4	87.5
HGAT w 2 head	87.2	86.8	88.4
HGAT w 3 head	**87.9**	**87.3**	**89.1**
HGAT w 4 head	87.7	87.2	88.7
HGAT w 5 head	87.5	86.9	88.2
Trans. w 5 head	86.6	86.1	88.4
Trans. w 6 head	87.6	86.7	88.7
Trans. w 7 head	87.5	87.0	89.1
Trans. w 8 head	**87.9**	**87.3**	**89.1**
Trans. w 9 head	86.8	87.1	88.8
Trans. w 10 head	87.2	87.3	89.0

Experiments

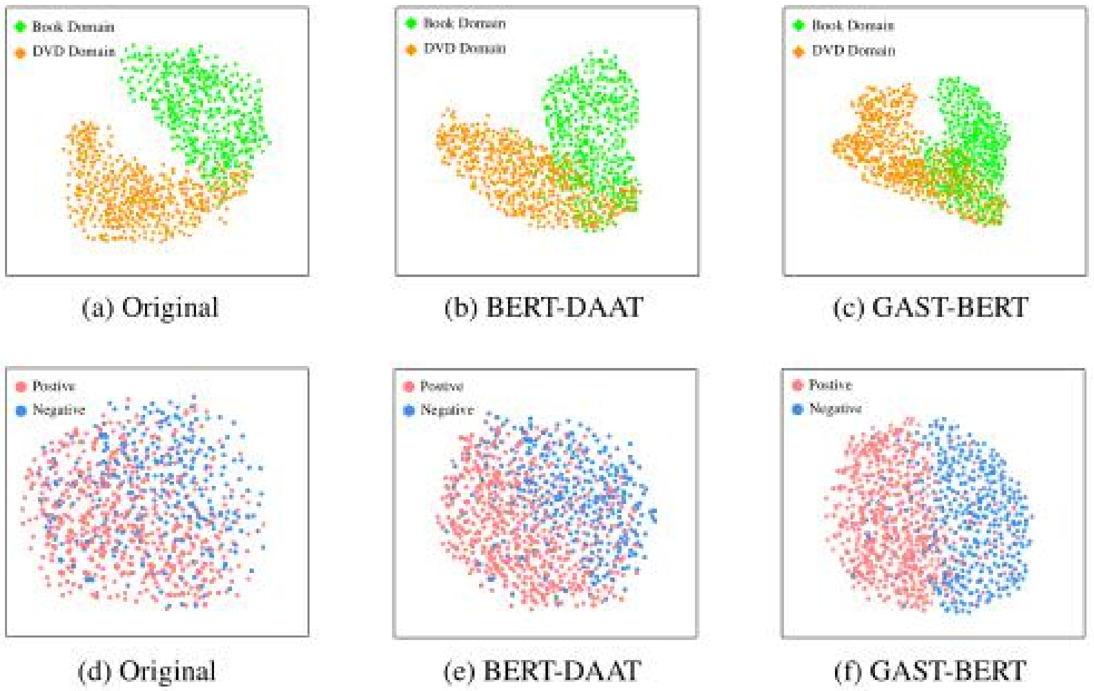


Figure 6: The t-SNE projection of the extracted features. The above three subfigures (i.e., (a)~(c)) show t-SNE visualization of different model's feature embedding for the $B \mapsto D$ task. The red and blue points in (d)~(f) denote the target positive and target negative examples, respectively.

Thank you！