Graph Adaptive Semantic Transfer for Cross-domain Sentiment Classification

Kai Zhang¹, Qi Liu¹, Zhenya Huang¹, Mingyue Cheng¹, Kun Zhang², Mengdi Zhang³, Wei Wu³, Enhong Chen¹

¹ University of Science and Technology of China
² Hefei University of Technology
³ Meituan

SIGIR 2022
1. Introduction
2. Related Work
3. The GAST Model
4. Experiment
5. Conclusions
Introduction

Cross-domain Sentiment Classification (CDSC)

Refers to the use the similarity knowledge learned in the source domain to the target domain.

- Source → Target
- Single → Invariant

Values

The use of similarities between tasks to apply knowledge learned from an old field to a new field.

- Unlabeled data
- Cold start
- Model versatility
- Weak computing ...

Source Domain $\sim P_S(X,Y)$

Target Domain $\sim P_T(Z,H)$

$D_S = \{(x_i,y_i), \forall i \in \{1,\ldots,N\}\}$

$D_T = \{(z_j,?), \forall j \in \{1,\ldots,M\}\}$

Unlabeled or limited labels

Lots of labeled data
Challenges

1. Making full use of domain shared features is pivotal for cross-domain sentiment classification. So, what information do we need to focus on in the whole sentence? How to pay attention?

2. How to mine and utilize this information that is critical for cross-domain classification tasks?
Related Work

Source Domain $\sim P_S(X, Y)$

\[ D_S = \{(x_i, y_i), \forall i \in \{1, \ldots, N\}\} \]

Target Domain $\sim P_T(Z, H)$

\[ D_T = \{(z_j, ?), \forall j \in \{1, \ldots, M\}\} \]

- Unlabeled problem
  - Sample re-weighting
  - Subspace matching
  - Deep methods

Maximum Mean Discrepancy M. Long, et al. ICML 2015

Reverse Gradient Y. Ganin and V. Lempitsky ICML 2015
Related Work

Domain-Adversarial Neural Networks (ICML 2015)

- **Input**: $x \in X$, input space of the image data
- **Output**: $y \in Y$, $Y = \{1, 2, 3, \ldots, n\}$, label space of image data
- **Goal**: predict the classification labels of input images in the target domain
IATN model

- Construct a better sentiment transfer network
- Make better use of Aspect information

Contains two parts:

1. S-Net
   - Learn shared information between domains

2. A-Net
   - Modeling Aspect Information
Limitation of previous methods

1. DNN-based/PLMs-based methods have become mainstream Encoders, and the model architecture is based on the evolution of the DANN framework
2. Design different deep, dedicated to mining better cross-domain semantic features

There are some problems
1. **Semantic representation:** Failure to fully utilize the internal information of the text (parsing & relation)
2. **Feature transfer:** ignoring important syntactic information in the process of cross-domain semantic transfer

GAST: Graph Adaptive Semantic Transfer for Cross-domain Sentiment Classification
1. Sentiment words play a crucial role in CDSC, POS-tags can distinguish sentiment words (e.g., “horrible” and “interesting” in the Figure) via the POS-tag “JJ” in a natural way, i.e., the “JJ” label means the word is an adjective.
1. The sentiment polarity of reviews is **largely influenced by** the sentiment word’s **neighbors**, whether they are in-domain or across-domain; **Meanwhile**, different neighbors’ syntactic relations also have different influences for each sentiment word.

---

**The GAST Model**

**Motivation of our GAST model**

2. **The sentiment polarity of reviews is largely influenced by the sentiment word’s neighbors**, whether they are in-domain or across-domain; **Meanwhile**, different neighbors’ syntactic relations also have different influences for each sentiment word.

---

**GAST: Graph Adaptive Semantic Transfer for Cross-domain Sentiment Classification**

This movie is unbelievably **horrible**, please reconsider watching.

The plot of this book was quite **interesting**; it is worth reading!
3. As shown in Figure, 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.
The GAST Model

Overall architecture of GAST

1. Generally, GAST improves the semantic representation and transferable knowledge between domains by aggregating the information from both word sequences and syntactic graphs.

2. GAST mainly contains two modules to learn comprehensive semantics:
   - The first is POS-based Transformer (POS-Transformer, a in Figure).
   - The other is Hybrid Graph Attention (HGAT, b in Figure).
The GAST Model

Detail of the GAST model

1. We project the word’s embedding matrix $E$ into the query, key, and value matrices
2. Apart from word’s embeddings, we also map the whole tag embedding matrix

\[ Z = \text{concat}(z_1, z_2, \ldots, z_I), \]
\[ z_i = \text{Att.}(Q, K, V_i) + \text{Att.}(Q^t, K^t, V_i), \]
\[ \text{Att.}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d/1}} \right) V, \]
The GAST Model

Detail of the GAST model

1. Contains two different calculation methods for better relation representation (i.e., relation-aggregation and relation-activation)

2. Through the above two relational functions, we can obtain two syntax-enhanced word representations

\[
\begin{align*}
    h_{agg_i}^{l+1} &= \| \kappa \sum_{k=1}^{K} \alpha_{ik} W_{lk} h_j^{l} \\
    f_{ij}^{lk} &= \sigma(\alpha_{ij}^{l}[W_{lk} h_i^{l}]W_{lk} h_j^{l} W_{lk} r_{ij}) \\
    \alpha_{ij}^{l} &= \exp\left(\frac{f_{ij}^{lk}}{\sum_{j=1}^{N_i} \exp(f_{ij}^{lk})}\right) \\
    \beta_{ij}^{l} &= \exp\left(\frac{F_{act}(h_i^{l}, h_j^{l})}{\sum_{j=1}^{N_i} \exp(F_{act}(h_i^{l}, h_j^{l}))}\right) \\
    F_{act} &= \frac{(W_{l} h_i^{l})(W_{k} h_j^{l} + W_{r} r_{ij})^T}{\sqrt{d_{i} d_{j} K}} \\
    h_{act_i}^{l+1} &= \| \kappa \sum_{j=1}^{K} \beta_{ij}^{l} (W_{l} h_i^{l} + W_{r} r_{ij}) \\
\end{align*}
\]
The GAST Model

Detail of the GAST model

**Integrated Adaptive Strategy**

As shown in Figure (c), the strategy includes three loss functions:

1. a **classifier loss** for sentiment knowledge learning
2. a **discriminator loss** for invariant feature extracting
3. a **feature alignment loss** for syntax-aware feature alignment

\[ L_c = - \frac{1}{n_s} \sum_{i=1}^{n_s} (y_s^i \ln \hat{y}_s^i + (1 - y_s^i) \ln(1 - \hat{y}_s^i)), \]

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

\[ L_a = - \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{C} \hat{y}_j^i \ln \hat{y}_j^i, \]
Experiment

1. We evaluate GAST on four widely-used Amazon datasets, i.e., DVD ($D$), Book ($B$), Electronics ($E$) and Kitchen ($K$)

2. We count the ratio of different syntactic relationships as shown in Figure. The proportions between various domains are close, meaning each sentence’s components might be remarkably similar, even in different domains

<table>
<thead>
<tr>
<th>Domains</th>
<th>Testing set percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Train</td>
</tr>
<tr>
<td>Books</td>
<td>1,600</td>
</tr>
<tr>
<td>DVD</td>
<td>1,600</td>
</tr>
<tr>
<td>Electronics</td>
<td>1,600</td>
</tr>
<tr>
<td>Kitchen</td>
<td>1,600</td>
</tr>
</tbody>
</table>


Our proposed model (BERT-GAST) gets further improvement and achieves a new SOTA. The reason is that GAST is able to consider the sequential information and syntactic structures jointly.
1. The vanilla transformer makes extra decisions on some unrelated word (e.g., “this”, “please”) and pays much attention to these uncritical words.

2. On the contrary, POS-transformer can alleviate this problem by revising attention scores with the help of POS tags.

3. HGAT could deal with the problem more appropriately through the domain-invariant syntactic relations between words.

---

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

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

---

(1) Transformer

(2) Pos-Transformer

(3) Hybrid GAT

---

The plot of this book was quite interesting; it is worth reading!
Experiment

Adaptive Efficiency & Adaptive Graph

The GAST model with 40% samples even performs better than IATN with 80% samples, which proves that GAST gains an advanced adaptive ability and efficiency in CDSC.

<table>
<thead>
<tr>
<th>Syntax Parser</th>
<th>$D \rightarrow B$</th>
<th>$D \rightarrow E$</th>
<th>$D \rightarrow K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Without Graph</td>
<td>86.6</td>
<td>85.9</td>
<td>88.1</td>
</tr>
<tr>
<td>(2) Stanford Graph</td>
<td>87.1 (+0.5)</td>
<td>86.6 (+0.7)</td>
<td>88.6 (+0.5)</td>
</tr>
<tr>
<td>+compare with (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Biaffine Graph</td>
<td>87.9 (+1.3)</td>
<td>87.3 (+1.4)</td>
<td>89.1 (+1.0)</td>
</tr>
<tr>
<td>+compare with (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The quality of the syntactic graph affect the final results, which shows the effectiveness of syntactic features for CDSC research.
Outline

1. Introduction
2. Related Work
3. The GAST Model
4. Experiment
5. Conclusions
Conclusion

In this Work

- Learn what?
  - Adaptive Graph Features (pos-tags & dependency relations)
- How to learn?
  - Graph Adaptive Semantic Transfer (GAST) model
- Experimental analysis

In the Feature

- Further characterization of graph structural information
- Dynamic semantic change & dynamic graph learning
See more details in our paper:

Graph Adaptive Semantic Transfer for Cross-domain Sentiment Classification