A Hierarchical Sequence-to-Set Model with Coverage Mechanism for Aspect Category Sentiment Analysis

Siyu Wang

Gusu Laboratory of Materials, Suzhou, China

2024.5.1

Introduction

Aspect-based sentiment analysis (ABSA) [Pontiki et al., 2014] is a fine-grained sentiment analysis task that involves many subtasks, aspect category detection (ACD) and aspect sentiment classification (ASC) are two of them.

Examples

"`The food is delicious, but the price is a bit expensive.`", the two aspect categories (`food, price`) are detected by ACD, and the sentiment polarities of detected aspect categories (`positive, negative`) can be predicted by ASC. In this paper, we focus on ACSA, which aims to jointly detect the discussed aspect categories (ACD) and their corresponding sentiment polarities (ASC) [Zhang et al., 2022]. For the previous example, ACSA models can directly predict two category-sentiment pairs (`food, positive`) and (`service, negative`).

Challenges

for ACSA, generative models still face three challenges:

- **(1)** How to alleviate the missing predictions, namely correctly predicting all category-sentiment pairs contained in a sentence.
- **(2)** Category-sentiment pairs are inherently a disordered set. Consequently, the model incurs a penalty even when its predictions are correct, but the predicted order is inconsistent with the ground truths.
- **(3)** It is crucial to ensure that different aspect categories focus on different sentiment words, and the polarity of the aspect category should be the aggregation of the polarities of these sentiment words.

Examples

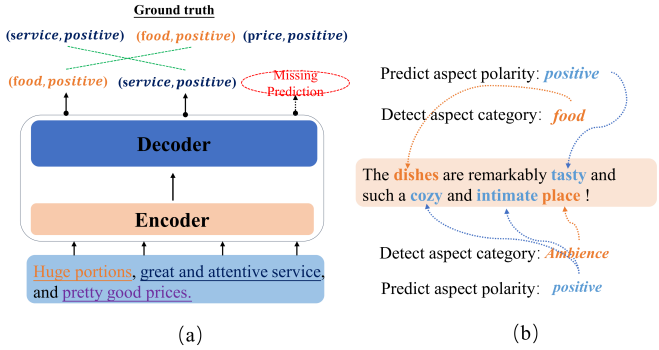


Figure 1: (a) An example of the missing prediction of aspect categories. (b) An example of different aspect categories focus on different sentiment words, where the final polarity is the aggregation of each polarity identified from the sentiment words.

Our Model

This paper proposes a hierarchical generative model with a coverage mechanism using sequence-to-set learning to tackle all three challenges simultaneously.

Problem Formalization

Given a predefined aspect category set $\mathcal{A} = \{a_1, a_2, ..., a_M\}$, sentiment polarity set $\mathcal{P} = \{positive, negative, neutral\}$, and a sentence $\mathbf{x}$ containing N words. Our task is to detect all the mentioned category-sentiment pairs $\mathbf{y}$ from $\mathbf{x}$, formulated as:

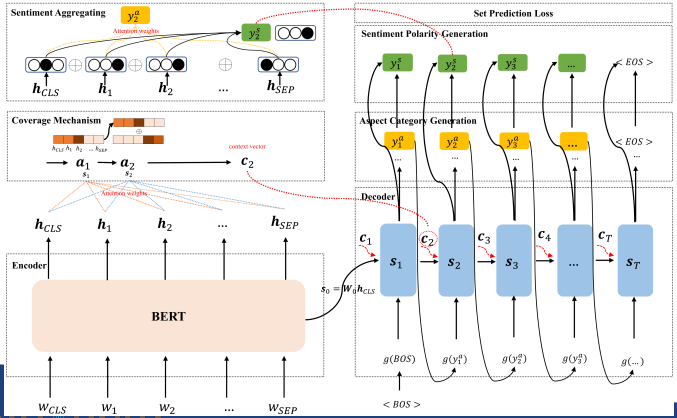
$$\mathbf{y} = [y_1, y_2, ..., y_T], \tag{1}$$

where $y_k = (y_k^a, y_k^s)$ is the k^{th} predicted aspect category and aspect sentiment polarity (category-sentiment pair). Consequently, the ACSA can be conceptualized as the search for an optimal sequence $\mathbf{y}$, which maximizes the conditional probability $p(\mathbf{y}|\mathbf{x})$. This probability is computed as:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{t=1}^{T} p(y_t^a|\mathbf{y}_{1:t-1}^a, \mathbf{x};\theta) p(y_t^s|y_t^a, \mathbf{x};\theta), \tag{2}$$

Model Architecture

An overview of our proposed model is shown in Figure. It consists of two parts: **Sentence Encoder** and **Decoder**.



Encoder

A sentence $\mathbf{x}$ in a review is composed of N words, which is formulated as:

$$\mathbf{x} = [w_1, w_2, ..., w_N], \tag{3}$$

where w_i denotes i^{th} word in the sentence. We employ BERT to encode $\mathbf{x}$ and output the context-aware representations $\mathbf{H} = [\mathbf{h}_{CLS}, \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N, \mathbf{h}_{SEP}]$. Then we adopt hidden state $\mathbf{h}_{CLS} \in \mathbb{R}^d$ to obtain the initial hidden state of the decoder, which is computed by:

$$\mathbf{s}_0 = \mathbf{W}_0 \mathbf{h}_{CLS} + \mathbf{b}_0, \tag{4}$$

Decoder

The probability of generating the t^{th} aspect category y_t^a is defined as:

$$p(y_t^a|\mathbf{y}_{1:t-1}^a, \mathbf{x}) = softmax(\mathbf{W}_1\mathbf{s}_t + b_1). \quad (5)$$

where $\mathbf{y}_{1:t-1}^a$ are previous generated aspect categories.

Coverage Mechanism

As mentioned above, a sentence usually contains one or more category-sentiment pairs. In order to alleviate the missing predictions, we introduce a coverage value, which can memorize the part covered by previous time steps in the sentence. According to the coverage value, the decoder will increase the attention weight for the words that have previously received less attention and decrease the attention weight for the words that have previously received more attention:

$$e_{t,j} = \mathbf{v}_a^T tanh(\mathbf{W}_a \mathbf{s}_{t-1} + \mathbf{U}_a \mathbf{h}_j + \mathbf{m}_a \tilde{c}_{t-1,j} + \mathbf{b}_a), \tag{6}$$

where $\mathbf{m}_a$ is the weight vector, and $\tilde{c}_{t-1,j}$ is the coverage value of word w_j at time step $t-1$ of the decoder, which is defined as:

$$\tilde{c}_{t-1,j} = \sum_{k=1}^{t-1} a_{k,j}. \tag{7}$$

Intuitively, $\tilde{c}_{t-1,j}$ denotes the degree of coverage derived by word w_j that has received the sum of attention weights at decoder time step $t-1$.

Hierarchical Generation Mechanism

Specifically, for time step t, the decoder firstly generates aspect category y_t^a by Equation (5). Then we compute the aspect-aware attention weight between the predicted aspect category y_t^a and source word w_j by:

$$a'_{t,j} = \frac{\exp\left(e'_{t,j}\right)}{\sum_{k=1}^{N+2} \exp\left(e'_{t,k}\right)}, \tag{8}$$

where $e'_{t,j}$ is computed by:

$$e'_{t,j} = \mathbf{v}_s^T tanh(\mathbf{W}_s g(y_t^a) + \mathbf{U}_s \mathbf{h}_j + \mathbf{b}_s). \tag{9}$$

Hierarchical Generation Mechanism

The polarity of an aspect should be the aggregation of the polarities of the sentiment words it emphasizes. Specifically, for word w_j, we predict its polarity by encoder's output $\mathbf{h}_j$:

$$\mathbf{p}_j = \mathbf{W}_{p2}(ReLU(\mathbf{W}_{p1}\mathbf{h}_j + \mathbf{b}_{p1}) + \mathbf{b}_{p2}), \tag{10}$$

where $\mathbf{p}_j \in \mathbb{R}^3$ represents the sentiment predictions of w_j belongs to $\{positive, negative, neutral\}$, respectively. Then we obtain the aspect category polarity by aggregating the word sentiment predictions based on the aspect-aware attention weight. For aspect category y_t^a, its probability of sentiment polarity is computed by:

$$\begin{aligned} p(y_t^s | y_t^a, \mathbf{x}) &= softmax(\theta \mathbf{p}_t^1 + (1-\theta)\mathbf{p}_t^2), \\ \mathbf{p}_t^1 &= \sum_{j=1}^{N+2} \mathbf{p}_j a_{t,j}^{'}, \\ \mathbf{p}_t^2 &= tanh(\mathbf{W}_{p3}\mathbf{s}_t + \mathbf{b}_{p3}), \end{aligned} \tag{11}$$

Model Optimization

The main difficulty of training is to score the predicted pairs with respect to the ground truths. In this scenario, it is not proper to apply the cross-entropy loss function to measure the difference between two sets, since cross-entropy loss is sensitive to the permutation of the predictions.

Model Optimization

we propose a set prediction loss that can produce an optimal bipartite matching between predicted and ground truth pairs. we first search for a permutation π^* with the lowest cost:

$$\pi^* = _{\pi \in \mathcal{O}_N} \sum_{i=1}^{N} C_{match}(y_i, p_{\pi(i)}), \tag{12}$$

where $\mathcal{O}_N$ is the space of all N-length permutations, and $C_{match}(.)$ is the matching cost function between ground truths and predicted pairs, which is computed by:

$$C_{match}(y_i, p_{\pi(i)}) = -\mathbb{I}_{y_i^a \neq \phi}[p^a_{\pi(i)}(y_i^a) + p^s_{\pi(i)}(y_i^s)], \tag{13}$$

where $p^a_{\pi(i)}, p^s_{\pi(i)}$ are aspect and sentiment probability distribution and computed by Equation(5,11), y_i^a, y_i^s are target aspect and sentiment, respectively. This optimal assignment π^* is computed efficiently by the Hungarian algorithm [Kuhn, 1955].

Model Optimization

The second step involves computing the loss function for all pairs identified in the preceding step. We define the loss function as follows:

$$\mathcal{L} = -\sum_{i=1}^{N}[\log p^{a}_{\pi^{*}(i)}(y^{a}_{i}) + \log p^{s}_{\pi^{*}(i)}(y^{s}_{i})]. \tag{14}$$

Datasets

We evaluate our model on four datasets, and the statistics of the datasets are shown in Table.

Dataset		#Pos	#Neg	#Neu	#Sen
MAMS	Train	2170	2343	3465	3549
	Test	245	263	393	400
Rest	Train	2177	839	500	2891
	Test	657	222	94	767
SRest	Test	379	136	80	595
MRest	Test	278	86	14	172

Table 1: Statistics of the experimental datasets. #Pos, #Neg, and #Neu mean the number of positive, negative, and neutral aspect categories on datasets, respectively. #Sen denotes the number of sentences on datasets.

Baselines

We compare our proposed model with classification and generative model baselines.

Category	Models	MAMS			Rest		
		P	R	F1	P	R	F1
Classsification	AddOneDim-LSTM †	58.742	59.563	59.150	72.349	71.326	71.834
	AddOneDim-TextCNN †	56.469	56.937	56.702	64.902	67.523	66.187
	AddOneDim-BERT †	73.246	71.809	72.520	79.943	80.301	80.122
	AS-Capsules	69.250	68.479	68.862	75.799	73.826	74.799
	MSS ⋆	-	-	-	82.520	77.040	79.680
	AC-MIMLLN-BERT	73.228	73.770	73.498	79.829	77.287	78.537
Generation	Seq2Seq-att †	58.239	54.384	56.245	69.696	62.076	65.666
	BART-ACSA †	72.888	71.624	72.250	81.979	80.884	81.428
Our	SGM	68.926	67.037	67.968	81.489	78.040	79.727
	CSGM	68.680	67.703	68.188	81.508	79.274	80.375
	HCSGM	73.882	73.437	73.659	81.438	81.295	81.366
	HCSGM+SL	74.922	73.363	**74.134**	81.578	82.049	**81.813**
Category	Models	SRest			MRest		
		P	R	F1	P	R	F1
Classsification	AddOneDim-LSTM †	66.284	72.325	69.173	85.057	69.753	76.649
	AddOneDim-TextCNN †	58.301	66.499	62.131	78.334	69.136	73.448
	AddOneDim-BERT †	75.488	82.073	78.643	88.652	77.513	82.709
	AS-Capsules	70.549	73.949	72.209	85.904	73.633	79.296
	AC-MIMLLN-BERT	73.166	76.303	74.701	90.735	76.896	83.244
Generation	Seq2Seq-att †	65.322	66.387	65.850	79.773	55.291	65.313
	BART-ACSA †	77.760	81.681	79.672	89.851	79.630	84.432
Our	SGM	78.162	80.392	79.261	87.897	74.339	80.551
	CSGM	78.103	81.120	79.583	87.924	76.367	81.739
	HCSGM	77.545	82.185	79.798	88.638	79.894	84.039
	HCSGM+SL	77.734	82.689	**80.135**	88.610	81.041	**84.657**

Table 2: Comparisons of baselines performances in ACSA. The evaluation results in terms of micro-Precision (P,%) micro-Recall (R,%) and micro-F1 (F1,%), and the baselines marked † are our implementations, ⋆ refers to citing from Shi et al. (2023). We run all models three times and report the average results on the test sets. The best F1 results are bold.

Case Study

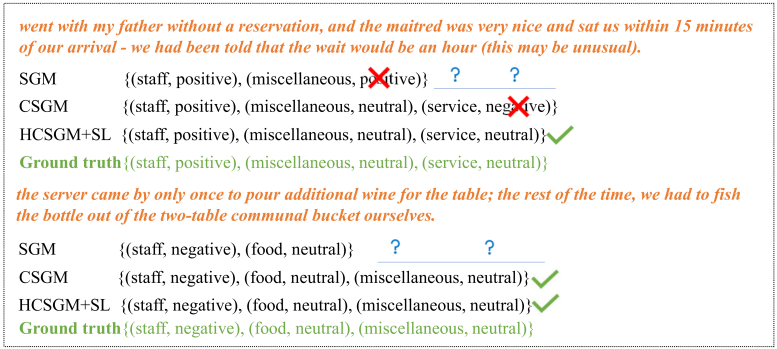


Figure 2: Case study on MAMS dataset. False prediction pairs are marked with "×" and missing pairs are marked with "?".

THANK YOU

References I

[Kuhn, 1955] Kuhn, H. W. (1955).
The hungarian method for the assignment problem.
Naval research logistics quarterly, 2(1-2):83–97.

[Pontiki et al., 2014] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014).
SemEval-2014 task 4: Aspect based sentiment analysis.
In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35. Association for Computational Linguistics.

[Zhang et al., 2022] Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2022).
A survey on aspect-based sentiment analysis: tasks, methods, and challenges.
IEEE Transactions on Knowledge and Data Engineering.