Deep Context- and Relation-Aware Learning for Aspect-based Sentiment Analysis

Shinhyeok Oh[1*†], **Dongyub Lee**[2*], **Taesun Whang**[3], **Ilnam Park**[4],
Gaeun Seo[4], **Eunggyun Kim**[4], **and Harksoo Kim**[5‡]
[1]Netmarble AI Center [2]Kakao Corp. [3]Wisenut Inc.
[4]Kakao Enterprise [5]Konkuk University

Abstract

Existing works for aspect-based sentiment analysis (ABSA) have adopted a unified approach, which allows the interactive relations among subtasks. However, we observe that these methods tend to predict polarities based on the literal meaning of aspect and opinion terms and mainly consider relations implicitly among subtasks at the word level. In addition, identifying multiple aspect–opinion pairs with their polarities is much more challenging. Therefore, a comprehensive understanding of contextual information w.r.t. the aspect and opinion are further required in ABSA. In this paper, we propose Deep Contextualized Relation-Aware Network (DCRAN), which allows interactive relations among subtasks with deep contextual information based on two modules (i.e., Aspect and Opinion Propagation and Explicit Self-Supervised Strategies). Especially, we design novel self-supervised strategies for ABSA, which have strengths in dealing with multiple aspects. Experimental results show that DCRAN significantly outperforms previous state-of-the-art methods by large margins on three widely used benchmarks.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a task of identifying the sentiment polarity of associated aspect terms in a sentence. Generally, ABSA is composed of three subtasks, 1) aspect term extraction (ATE), 2) opinion term extraction (OTE), and 3) aspect-based sentiment classification (ASC). Given the sentence “*Food is good, but service is dreadful.*”, ATE aims to identify two-aspect terms “*food*” and “*service*”, and OTE aims to determine two-opinion terms “*good*” and “*dreadful*”. Then, ASC assigns a sentiment polarity of each aspect: “*food* (*positive*)” and “*service* (*negative*)”.

	Examples (Ground Truth)	Model	Aspect (Polarity)	Opinion
E1	I’ve had *better Japanese food* (neg) at a mall food court.	RACL	Japanese food (pos)	better
		DCRAN	Japanese food (neg)	better
E2	The *sushi* (neg) is cut in blocks *bigger* than my cell phone.	RACL	sushi (neu)	bigger
		DCRAN	sushi (neg)	bigger
E3	While the *smoothies* (neg) are a little *bigger* for me, the *fresh juices* (pos) are the *best* I have ever had !	RACL	smoothies (pos) fresh juices (pos)	bigger fresh best
		DCRAN	smoothies (neg) fresh juices (pos)	bigger best

Table 1: Examples of ABSA results comparing to previous approach (Chen and Qian, 2020) that we reimplement. All the results are based on BERT$_{base}$ model for a fair comparison. The polarity labels pos, neu, and neg, denote positive, neutral, and negative, respectively.

Existing works for ABSA have adopted a two-step approach, which considers each subtask separately (Tang et al., 2016; Xu et al., 2018). However, most recently, unified approaches have achieved significant performance improvements in ABSA task. Luo et al. (2020) focused on modeling the interactions between aspect terms and Chen and Qian (2020) exploited dyadic and triadic relations between subtasks (i.e., ATE, OTE, ASC).

Despite the impressive results, their methods have two limitations. First, they only consider relations among subtasks at the word level and do not explicitly utilize contextualized information of the whole sequence. For example, E1 in Table 1, the opinion term “*better*” seems to represent positive opinion of “*Japanese food*”. However, the authentic meaning of E1 is “*The Japanese food I have had at the food court was more delicious than the one I had at this restaurant*”. Thus, previous approaches tend to assign polarities based on the literal meaning of aspect and opinion terms (E2). Second, identifying multiple aspect–opinion pairs and their polarities is much more challenging as the model needs to not only detect multiple aspects and

[*] These two authors equally contributed to this work.

[†] This work was done while the author was an intern at Kakao Enterprise.

[‡] Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 495–503
August 1–6, 2021. ©2021 Association for Computational Linguistics

opinions but also correctly predict each polarity of the aspect (E3).

To address the aforementioned issues, we propose Deep Contextualized Relation-Aware Network (DCRAN) for ABSA. DCRAN not only implicitly allows interactive relations among the subtasks of ABSA, but also explicitly considers their relations by using contextual information. Our main contributions are as follows: 1) We design aspect and opinion propagation decoder so that the model has a comprehensive understanding of the whole context, and thus it results in better prediction of the polarity. 2) We propose novel self-supervised strategies for ABSA, which are highly effective in dealing with multiple aspects and considering deep contextualized information with the aspect and opinion terms. To the best of our knowledge, it is the first attempt to design explicit self-supervised methods for ABSA. 3) Experimental results demonstrate that DCRAN significantly outperforms previous state-of-the-art methods on three widely used benchmarks.

2 DCRAN: Deep Contextualized Relation-Aware Network

2.1 Task Definition

Given a sentence $S = \{w_1, w_2, ..., w_n\}$, where n denotes the number of tokens, we aim to solve three subtasks: aspect term extraction (ATE), opinion term extraction (OTE), and aspect-based sentiment classification (ASC) as sequence labeling problems. ATE task aims to identify a sequence of aspect terms $Y^a = \{y_1^a, y_2^a, ..., y_n^a\}$, where $y_i^a \in \{B, I, O\}$, and OTE task aims to identify a sequence of opinion terms $Y^o = \{y_1^o, y_2^o, ..., y_n^o\}$, where $y_i^o \in \{B, I, O\}$ of aspect and opinion terms, respectively. Likewise, ASC task aims to assign a sequence of polarities $Y^p = \{y_1^p, y_2^p, ..., y_n^p\}$, where $y_i^p \in \{POS, NEU, NEG, O\}$. The labels *POS*, *NEU*, and *NEG* denote *positive*, *neutral*, and *negative*, respectively.

2.2 Task-Shared Representation Learning

Following existing works, we utilize pre-trained language models, such as BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) as the shared encoder to construct context representation, which is shared by subtasks: ATE, OTE, and ASC. Given a sentence $S = \{w_1, w_2, ..., w_n\}$, pre-trained language models take the input sequence, $\mathbf{X}_{\text{absa}} = [[\texttt{CLS}]\, w_1\, w_2 \dots w_n\, [\texttt{SEP}]]$, and output a sequence of the shared context representation, $H = \{h_{[\text{CLS}]}, h_1, h_2, ..., h_n, h_{[\text{SEP}]}\} \in \mathbb{R}^{d_h \times (n+2)}$, where d_h represents a dimension of the shared encoder. We represent the parameters of the shared encoder as Θ_s. Then, we utilize a single-layer feed-forward neural network (FFNN) as,

$$\begin{aligned} Z^a &= (W_1 h_{[1:n+1]} + b_1) \\ \hat{Y}^a &= \text{softmax}(W_2 Z^a + b_2), \end{aligned} \tag{1}$$

where $W_1 \in \mathbb{R}^{d_h \times d_h}$ and $W_2 \in \mathbb{R}^{3 \times d_h}$ are trainable parameters. The parameters of a single-layer FFNN are represented as Θ_a for aspect term extraction. The objective of aspect term extraction is minimizing the negative log-likelihood (NLL) loss: $\mathcal{L}_{\text{ate}}(\Theta_s, \Theta_a) = -\sum \log p(Y^a|H)$. Likewise, Z^o and $\hat{Y}^o$ are obtained as in Equation 1. Then, the NLL loss of opinion term extraction is defined as, $\mathcal{L}_{\text{ote}}(\Theta_s, \Theta_o) = -\sum \log p(Y^o|H)$.

2.3 Aspect and Opinion Propagation

We utilize the transformer-decoder (Vaswani et al., 2017) to consider relations of aspect and opinion while predicting a sequence of polarities. Our transformer-decoder is mainly composed of a multi-head self-attention, two multi-head cross attention, and a feed-forward layer. The multi-head self-attention takes shared context representation H as,

$$U^h = \text{LN}(H + \text{SelfAttn}(H, H, H)) \tag{2}$$

and U^h, Z^a, and Z^o are fed into two steps of cross multi-head attention as,

$$U^a = \text{LN}(U^h + \text{CrossAttn}(U^h, Z^a, Z^a)) \tag{3}$$

$$U^o = \text{LN}(U^a + \text{CrossAttn}(U^a, Z^o, Z^o)) \tag{4}$$

where LN represents layer norm (Ba et al., 2016). Note that Equation 3 and 4 represent aspect and opinion propagation, respectively. Then U^o is fed into a single-layer FFNN to obtain a sequence of polarities Y^p. The objective of aspect-based sentiment analysis is minimizing the NLL loss: $\mathcal{L}_{\text{asc}}(\Theta_s, \Theta_a, \Theta_o, \Theta_p) = -\sum \log p(Y^p|H, Z^a, Z^o)$. The architecture of the aspect and opinion propagation is described in Figure 1-(a).

2.4 Explicit Self-Supervised Strategies

To further exploit the aspect–opinion relation with contextualized information of a sentence, we propose explicit self-supervised strategies consisting of two auxiliary tasks: 1) type-specific masked term

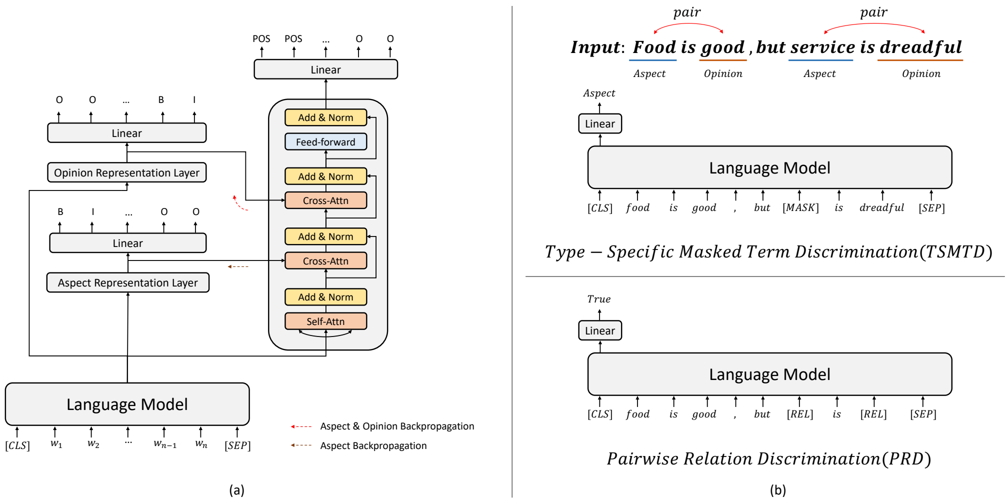
POS
POS
...
O
O
Linear
Add & Norm
Feed-forward
Add & Norm
Cross-Attn
Add & Norm
Cross-Attn
Add & Norm
Self-Attn
O
O
...
B
I
Linear
Opinion Representation Layer
B
I
...
O
O
Linear
Aspect Representation Layer
Language Model
[CLS]
w_1
w_2
...
w_{n-1}
w_n
[SEP]
Aspect & Opinion Backpropagation
Aspect Backpropagation
(a)
pair
pair
Input: Food is good , but service is dreadful
Aspect
Opinion
Aspect
Opinion
Aspect
Linear
Language Model
[CLS] food is good , but [MASK] is dreadful [SEP]
Type − Specific Masked Term Discrimination(TSMTD)
True
Linear
Language Model
[CLS] food is good , but [REL] is [REL] [SEP]
Pairwise Relation Discrimination(PRD)
(b)

Figure 1: Overall architecture of Deep Contextualized Relation-Aware Network (DCRAN) for ABSA.

discrimination (TSMTD) and 2) pairwise relations discrimination (PRD). The examples of *Explicit Self-Supervised Strategies* are described in Figure 1-(b).

Type-Specific Masked Term Discrimination In the type-specific masked term discrimination task, we uniformly mask aspects, opinions, and terms that do not correspond to both, using the special token [MASK]. The input sequence of a masked sentence is represented as, $\mathbf{X}_{\text{tsmtd}} = [[\text{CLS}]\, w_1 \ldots [\text{MASK}]_i \ldots w_n\, [\text{SEP}]]$, and is fed into pre-trained language models. Then, the output representation of [CLS] token is used to classify which type of term is masked in a sentence as,

$$\hat{Y}^m = \text{softmax}(W_3 h_{[\text{CLS}]} + b_3),$$

where $W_3 \in \mathbb{R}^{3 \times d_h}$ represents trainable parameters and $\hat{Y}^m \in \{Aspect, Opinion, O\}$. The parameters of a linear projection layer are represented as Θ_m for the type-specific masked term discrimination. Then, the NLL loss of the type-specific masked term discrimination is defined as: $\mathcal{L}_{\text{tsmtd}}(\Theta_s, \Theta_m) = -\sum \log p(Y^m|H)$. This allows the model to explicitly exploit sentence information by discriminating what kind of term is masked.

Pairwise Relations Discrimination In this task, we uniformly replace both aspects and opinion terms using the special token [REL]. The input sequence of a masked sentence is represented as, $\mathbf{X}_{\text{prd}} = [[\text{CLS}]\, w_1 \ldots [\text{REL}]_i \ldots [\text{REL}]_j \ldots w_n\, [\text{SEP}]]$, and is fed into pre-trained language models. Then, the output representation of [CLS] token is used to discriminate whether the replaced tokens have a pairwise relation as,

$$\hat{Y}^r = \text{softmax}(W_4 h_{[\text{CLS}]} + b_4),$$

where $W_4 \in \mathbb{R}^{2 \times d_h}$ represents trainable parameters and $\hat{Y}^r \in \{True, False\}$. The parameters of a linear projection layer are represented as Θ_r for the pairwise relations discrimination. Then, the NLL loss of the pairwise relations discrimination is defined as: $\mathcal{L}_{\text{prd}}(\Theta_s, \Theta_r) = -\sum \log p(Y^r|H)$. We describe the negative sampling method to replace aspects and opinion terms in Appendix A.2.

2.5 Joint Learning Procedure

All these tasks are jointly trained, and the final objective is defined as,

$$\begin{aligned}
\mathcal{L}_{\text{absa}} &= \mathcal{L}_{\text{ate}} + \mathcal{L}_{\text{ote}} + \mathcal{L}_{\text{asc}} \\
\mathcal{L}_{\text{aux}} &= \mathcal{L}_{\text{tsmtd}} + \mathcal{L}_{\text{prd}} \\
\mathcal{L}_{\text{final}} &= \mathcal{L}_{\text{absa}} + \alpha \mathcal{L}_{\text{aux}}
\end{aligned}$$

where α is a hyper-parameter determining the degree of auxiliary tasks. Note that the parameters Θ_s are optimized for all subtasks. Especially, the parameters Θ_s are further optimized through $\mathcal{L}_{\text{tsmtd}}$ and $\mathcal{L}_{\text{prd}}$ to explicitly exploit the relations between aspect and opinion with context meaning.

3 Experiments

3.1 Experimental Setup

We evaluate our model on three widely used sentiment analysis benchmarks: laptop reviews (LAP14), restaurant reviews (REST14) from (Pontiki et al., 2014), and restaurant reviews (REST15)

		LAP14				REST14				REST15			
		ATE-F1	OTE-F1	ASC-F1	ABSA-F1	ATE-F1	OTE-F1	ASC-F1	ABSA-F1	ATE-F1	OTE-F1	ASC-F1	ABSA-F1
MNN (Wang et al., 2018)	GloVe	76.94	77.77	65.98	53.80	83.05	84.55	68.45	63.87	70.24	69.38	57.90	56.57
E2E-TBSA (Li et al., 2019)	GloVe	77.34	76.62	68.24	55.88	83.92	84.97	68.38	66.60	69.40	71.43	58.81	57.38
DOER (Luo et al., 2019)	GloVe	80.21	-	60.18	56.71	84.63	-	64.50	68.55	67.47	-	36.76	50.31
IMN^{-d} (He et al., 2019)	GloVe	78.46	78.14	69.62	57.66	84.01	85.64	71.90	68.32	69.80	72.11	60.65	57.91
RACL (Chen and Qian, 2020)	GloVe	81.99	79.76	71.09	60.63	85.37	85.32	74.46	70.67	72.82	78.06	68.69	60.31
WHW (Peng et al., 2020)	GloVe	-	74.84	-	62.34	-	82.45	-	71.95	-	78.02	-	65.79
IKTN (Liang et al., 2020)	BERT$_{base}$	80.89	78.90	73.42	62.34	86.13	86.62	74.35	71.75	71.63	76.79	69.85	62.33
SPAN (Hu et al., 2019)	BERT$_{large}$	82.34	-	62.50	61.25	86.71	-	71.75	73.68	74.63	-	50.28	62.29
IMN^{-d} (He et al., 2019)	BERT$_{large}$	77.55	**81.00**	75.56	61.73	84.06	85.10	75.67	70.72	69.90	73.29	70.10	60.22
Dual-MRC (Mao et al., 2021)	BERT$_{large}$	82.51	-	75.97	65.94	86.60	-	**82.04**	75.95	**75.08**	-	73.59	65.08
RACL (Chen and Qian, 2020)	BERT$_{large}$	81.79	79.72	73.91	63.40	86.38	**87.18**	81.61	75.42	73.99	76.00	74.91	66.05
DCRAN (Ours)	BERT$_{base}$	81.76	78.84	77.02	65.18	88.21	86.36	78.67	75.77	71.61	75.86	73.30	63.19
	BERT$_{large}$	**83.40**	79.72	**78.75**	**68.07**	**88.73**	86.07	80.64	**77.28**	74.45	**78.45**	**76.30**	**67.92**
	ELECTRA$_{base}$	**85.69**	**80.19**	79.36	70.22	89.64	87.30	84.12	80.00	77.41	78.80	**78.55**	71.67
	ELECTRA$_{large}$	85.61	79.77	**80.78**	**71.47**	**89.67**	**87.59**	**84.22**	**80.32**	**79.68**	**79.90**	77.99	**73.67**

Table 2: Evaluation results on the LAP14, REST14, and REST15 datasets, which are provided by Chen and Qian (2020). All the results except ours are cited from the existing works (Chen and Qian, 2020; Peng et al., 2020; Mao et al., 2021) and all the baselines are described in Appendix A.4. We report average results over five runs with random initialization. The best scores are in bold, and the second-best scores are underlined depending on the types of the pre-trained language model. '-' denotes unreported results.

from (Pontiki et al., 2015). Primitive versions of these benchmarks only provide aspect terms and sentiment polarities, while opinion terms are provided by Wang et al. (2016, 2017) later. Recently, Fan et al. (2019) provides aspect-opinion pairwise datasets (Section 2.4). Following He et al. (2019), we set four evaluation metrics: ATE-F1, OTE-F1, ASC-F1, and ABSA-F1. The ATE-F1, OTE-F1, and ASC-F1 measure each subtask's F-1 scores, and ABSA-F1 measures complete ABSA, which counts only when both ATE and ASC predictions are correct.

3.2 Quantitative Results

Table 2 reports the quantitative results on the LAP14, REST14, and REST15 datasets. Our experiments utilize two pre-trained language models such as BERT and ELECTRA, for the shared encoder. First, we observe that DCRAN-BERT$_{base}$ shows slightly lower ABSA-F1 scores than previous state-of-the-art methods, which is based on BERT$_{large}$, on the REST14 and LAP14 datasets except for the REST15 dataset. This suggests that our proposed methods are highly effective for ABSA. Overall, DCRAN-BERT$_{large}$ significantly outperforms previous state-of-the-art methods in all metrics. Another observation is that ELECTRA based models outperform BERT based models. As a result, DCRAN-ELECTRA$_{large}$ achieves absolute gains over previous state-of-the-art results by 5.5%, 4.4%, and 7.6% in ABSA-F1 on the LAP14, REST14, and REST15 datasets, respectively.

3.3 Ablation Study

To study the effectiveness of the aspect propagation (AP), opinion propagation (OP), type-specific masked term discrimination (TSMTD), and pairwise relations discrimination (PRD), we conduct ablation experiments on the REST14 dataset. We set the baseline model that did not utilize aspect and opinion propagation and explicit self-supervised strategies. When the AP and OP are not utilized, a single-layer FFNN is utilized as in Equation 1 to predict a sequence of polarities Y^p instead of transformer-decoder. As shown in Table 3, we can observe that the AP is more effective than the OP, and scores drop significantly when not utilizing the AP and OP. In the case of explicit self-supervised strategies, we can observe that the PRD is more effective than the TSMTD. As the PRD objective is discriminating whether the replace tokens have a pairwise aspect–opinion relations, it allows the model to more exploit the relations between aspect and opinion at a sentence level.

		ABSA-F1
	DCRAN-ELECTRA$_{base}$	**80.00**†
Aspect and Opinion Propagation	w/o AP	79.44†
	w/o OP	79.58†
	w/o AP & OP	79.08†
Explicit Self-Supervised Strategies	w/o TSMTD	79.56†
	w/o PRD	79.40†
	w/o TSMTD & PRD	79.03†
Baseline	w/o & AP & OP & TSMTD & PRD	78.61

Table 3: Ablation study on the REST14 dataset. We choose DCRAN-ELECTRA$_{base}$ as the baseline. † denotes statistical significance (p-value $<$ 0.05).

3.4 Aspect Analysis

We conduct aspect analysis by comparing sentences with single- and multiple-aspect. As shown in Table 4, *Aspect and Opinion Propagation* signif-

		REST14		REST15	
		ABSA-F1	Sent-level Acc.	ABSA-F1	Sent-level Acc.
Single-Aspect	DCRAN_ELECTRA$_{base}$	**78.62**	**74.48**	**66.23**	**67.69**
	w/o TSMTD & PRD	78.42	73.79	64.21	66.67
	w/o TSMTD & PRD & AP & OP	77.45	73.10	62.50	64.29
Multiple-Aspect	DCRAN_ELECTRA$_{base}$	**81.19**	**64.24**	**68.20**	**52.34**
	w/o TSMTD & PRD	80.22	61.70	65.16	48.60
	w/o TSMTD & PRD & AP & OP	79.88	61.39	64.84	46.73

Table 4: Aspect analysis on the REST14 and REST15 datasets. Comparisons of ABSA-F1 and sentence-level accuracy results for the case when the sentence contains single-aspect or multiple-aspect.

icantly improves performance when the sentence contains a single-aspect, while a small increase is observed w.r.t. the case of multiple-aspect. Although considering the relations between aspect and opinion implicitly can improve performance w.r.t. the case of single-aspect, it is not sufficient for inducing performance improvement for the multiple-aspect case. It suggests that additional explicit tasks are further required to identify multiple-aspect with corresponding opinions, which helps the model assign polarities correctly. In the case of multiple-aspect, *Explicit Self-Supervised Strategies* show absolute ABSA-F1 improvements of 0.97% (80.22% $\rightarrow$ 81.19%) and 3.04% (65.16% $\rightarrow$ 68.20) on the REST14 and REST15 datasets, respectively. This indicates explicit self-supervised strategies are highly effective for correctly identifying ABSA when the sentence contains multiple-aspect. In addition, the performance gain by *Explicit Self-Supervised Strategies* in Table 3 is mostly derived from the multiple-aspect cases (+0.97%), thus our proposed model has strengths in dealing with multiple aspects.

In ABSA, it is important to accurately predict all aspects and corresponding sentiment polarities in one sentence. Since ABSA-F1 is a word-level based metric, it still has a limitation to evaluate whether all aspects and corresponding polarities are correct or not. Therefore, we also evaluate our method with sentence-level accuracy; the number of sentences that accurately predicted all aspects and polarity in a sentence divided by total number of sentences. Unlike ABSA-F1, the sentence-level accuracy of multiple-aspect is lower than that of single-aspect, which implies identifying multiple aspects and their polarities is more challenging. In the case of multiple-aspect, our *Explicit Self-Supervised Strategies* leads significant sentence-level accuracy improvements of 2.54% (61.70% $\rightarrow$ 64.24%) and 3.74% (48.60% $\rightarrow$ 52.34%) on the REST14 and REST15 datasets, respectively. However, we observe only small improvements in sentence-level accuracy on both datasets when the sentence contains single-aspect. From these observations, we demonstrate that our proposed method is highly effective for the case when the sentence contains multiple aspects.

4 Conclusion

In this paper, we proposed the Deep Contextualized Relation-Aware Network (DCRAN) for aspect-based sentiment analysis. DCRAN allows interaction between subtasks implicitly in a more effective manner and two explicit self-supervised strategies for deep context- and relation-aware learning. We obtained the new state-of-the-art results on three widely used benchmarks.

Acknowledgements

We thank the anonymous reviewers, Dongsuk Oh, Jungwoo Lim, and Heuiseok Lim for their constructive comments. This work was partially supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1F1A1069737).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 3685–3694.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2vlda: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, 91:127–137.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515.

Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. 2021. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Hao Li and Wei Lu. 2017. Learning latent sentiment scopes for entity-level sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4194–4200.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. An iterative knowledge transfer network with routing for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.01935*.

Huaishao Luo, Lei Ji, Tianrui Li, Daxin Jiang, and Nan Duan. 2020. Grace: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 54–64.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: Dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4737–4742.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8600–8607.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Feixiang Wang, Man Lan, and Wenting Wang. 2018. Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 3316–3322.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598.

Jianfei Yu, Jing Jiang, and Rui Xia. 2018. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):168–177.

A Appendix

A.1 Related Work

Existing works have studied a two-step approach for ABSA. In a two-step approach, each model for ATE, OTE, and ASC are separately trained and are merged in a pipelined manner (Wang et al., 2016; Tang et al., 2016; Wang et al., 2017; He et al., 2017; Xu et al., 2018; Yu et al., 2018; Li et al., 2018; Chen and Qian, 2019). However, the errors from other tasks can be propagated to the ASC and can degrade performance after all.

Most recently, a unified approach that comprised of joint approach (García-Pablos et al., 2018; Luo et al., 2019; He et al., 2019; Luo et al., 2020) and collapsed approach (Li and Lu, 2017; Ma et al., 2018; Wang et al., 2018; Li et al., 2019) has been proposed. A joint approach labels each word with different tag sets for each task: ATE, OTE, and ASC. On the other hand, a collapsed approach labels each word as the combined one of ATE and ASC, such as "B-POSITIVE" and "I-POSITIVE", where "B" and "I" represent the aspect term boundary, and "POSITIVE" represents polarity. However, in a collapsed approach, the relations among subtasks cannot be effectively exploited because subtasks need to share all representation without distinction of each task. Therefore, a joint training approach allows the interactive relations between subtasks, while a collapsed approach is not.

A.2 Negative Sampling Algorithm for Pairwise Relations Discrimination

Algorithm 1 describes the negative sampling procedure in pairwise relations discrimination. The get_sample function takes a list of aspect-opinion pairs in a sentence and replaces them with [REL] tokens. Then, if the replaced tokens have pairwise relations, set the target label as True, and set as False if not. The get_pair function randomly selects a pairwise aspect and opinion, and the get_neg_pair function selects aspects and opinions of different pairs when there are two or more pairs in a sentence.

A.3 Implementation Details

We implemented our model by using the PyTorch (Paszke et al., 2019) deep learning library based on the open source[1] (i.e., Transformers (Wolf et al., 2020)). For the shared encoder, we adopt four types of pre-trained language models: BERT$_{base}$, BERT$_{large}$, ELECTRA$_{base}$, and ELECTRA$_{large}$. We set the batch size to 64 for the $base$ model, 12 for the BERT$_{large}$ and 32 for the ELECTRA$_{large}$. We set the initial learning rate to 5e-5 for BERT$_{base}$ and ELECTRA$_{base}$, 2e-5 for BERT$_{large}$, and 5e-6 for ELECTRA$_{large}$. For the transformer decoder, we set the number of heads in multi-head attention and hidden layers to 2 among range from 2 to 6, and hidden dimension size to 768. In the case of α, we obtained the best results when α is 1. The average runtime for each approach was about 20 seconds for BERT$_{base}$ and ELECTRA$_{base}$, and 90 seconds for BERT$_{large}$ and ELECTRA$_{large}$. We train our models using AdamP (Heo et al., 2021) optimizer and conduct experiments with Tesla V100 GPU for all the experiments.

Algorithm 1 Negative Sampling Algorithm for Pairwise Relations Discrimination

```
Input: $pairs$: list of aspect–opinion pairs in a sentence
Output: $pair$, $target$
  function GET_SAMPLE($pairs$)
      if count($pairs$) == 0 then
          return None, None
      else if count($pairs$) == 1 then
          return $pairs$[0], True
      else
          $random = \{0 < random \leq 1\}$
          if $random \leq 0.25$ then
              return get_pair($pairs$), True
          else
              return get_neg_pair($pairs$), False
```

A.4 Baselines

We compare our model with the following previous works[2].

MNN (Wang et al., 2018) is a multi-task model for ATE and ASC using attention mechanisms to learn the joint representation of aspect and polarity relations.

E2E-TBSA (Li et al., 2019) is an end-to-end model of the collapsed approach for ATE and ASC. Additionally, it introduces the auxiliary OTE task without explicit interaction.

[1] https://github.com/huggingface/transformers

[2] We do not compare our work with GRACE (Luo et al., 2020) as Luo et al. (2020) contains *conflict* tag in polarities.

	Examples (Ground Truth)	Model	Aspect (Polarity)	Opinion
E1	I have worked in restaurants and cook a lot, and there is no way a maggot should be able to get into *well prepared* *food* (neg).	RACL	food (pos)	well
		DCRAN w/o	food (pos)	well prepared
		DCRAN	food (neg)	well prepared
E2	All in all, I would return - as it was a *beautiful* *restaurant* (pos) - but I hope the *staff* (neg) pays more attention to the little details in the future.	RACL	-	-
		DCRAN w/o	restaurant (pos) staff (pos)	beautiful
		DCRAN	restaurant (pos) staff (neg)	beautiful
E3	I have never been so *disgusted* by both *food* (neg) and *service* (neg)	RACL	food (pos) service (pos)	disgusted
		DCRAN w/o	food (pos) service (neg)	disgusted
		DCRAN	food (neg) service (neg)	disgusted

Table 5: Case study on the REST15 dataset. Model comparison between previous state-of-the-art method (RACL) (Chen and Qian, 2020) and our proposed method (DCRAN). DCRAN w/o denotes DCRAN without *Explicit Self-Supervised Strategies* (Section 2.4). All models are built based on the BERT$_{base}$ model. The polarity labels pos, neu, and neg denote positive, neutral, and negative, respectively. '-' denotes that the model failed to extract corresponding terms.

DOER (Luo et al., 2019) is a dual cross-shared RNN framework that jointly trains ATE and ASC. It considers relations between aspect and polarity.

IMN (He et al., 2019) is a multi-task model for ATE and ASC with separate labels. The OTE task is fused into ATE by constructing five-class labels.

WHW (Peng et al., 2020) is a unified two-stage framework to extract (aspect, opinion, polarity) triples as a result of ATE, OTE, and ASC.

IKTN (Liang et al., 2020) is an iterative knowledge transfer network for ABSA considering the semantic correlations among the ATE, OTE, and ASC.

SPAN (Hu et al., 2019) is a pipeline approach to solve ATE and ASC using BERT$_{large}$. It uses a multi-target extractor for ATE and a polarity classifier for ASC.

RACL (Chen and Qian, 2020) defines interactive relations among ATE, OTE, and ASC. It proposes relation propagation mechanisms through the stacked multi-layer network.

Dual-MRC (Mao et al., 2021) leverages two machine reading comprehension problems to solve ATE and ASC. It jointly trains two BERT-MRC models sharing parameters.

A.5 Case Study

In E1 and E3, while all models correctly extract both aspect and opinion, RACL and DCRAN w/o make inaccurate polarities predictions based on the words having superficial meaning (i.e., *well prepared*, *disgusted*). Especially, E3 expresses a sarcastic opinion about aspect terms throughout the sentence. It suggests that these models cannot understand the authentic meaning of the sentence. On the other hand, DCRAN grasps the entire context and predicts the correct polarity corresponding to its aspect. In E2, the evidence for understanding the actual meaning of the aspect term *staff* is not specified in a word-level opinion and expressed in a sentence like "*I hope the staff pays more attention to the little details in the future*". In this case, RACL can not extract aspect and opinion terms, and DCRAN w/o make inaccurate polarities predictions for the aspect term *staff* based on the opinion term *beautiful*. However, DCRAN with *Explicit Self-Supervised Strategies* understands the sentence expressing an opinion on the *staff* and predicts correctly.