USSA: A Unified Table Filling Scheme for Structured Sentiment Analysis

Zepeng Zhai[1], **Hao Chen**[2], **Ruifan Li**[1,3,4]*, **Xiaojie Wang**[1,3,4]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
[2]STCA, Microsoft, China
[3]Engineering Research Center of Information Networks, Ministry of Education, China
[4]Key Laboratory of Interactive Technology and Experience System,
Ministry of Culture and Tourism, China
{zepeng, rfli, xjwang}@bupt.edu.cn and hche@microsoft.com

Abstract

Most previous studies on Structured Sentiment Analysis (SSA) have cast it as a problem of bi-lexical dependency parsing, which cannot address issues of overlap and discontinuity simultaneously. In this paper, we propose a niche-targeting and effective solution. Our approach involves creating a novel bi-lexical dependency parsing graph, which is then converted to a unified 2D table-filling scheme, namely USSA. The proposed scheme resolves the kernel bottleneck of previous SSA methods by utilizing 13 different types of relations. In addition, to closely collaborate with the USSA scheme, we have developed a model that includes a proposed bi-axial attention module to effectively capture the correlations among relations in the rows and columns of the table. Extensive experimental results on benchmark datasets demonstrate the effectiveness and robustness of our proposed framework, outperforming state-of-the-art methods consistently[1].

1 Introduction

Structured Sentiment Analysis (SSA) aims to identify all opinion tuples within a given sentence. An opinion tuple (***h***, ***t***, ***e***, ***p***) denotes a group of four elements: the holder ***h*** expresses a sentiment polarity ***p*** towards an opinion target ***t*** through a sentiment expression ***e***. As shown in Figure 1(a), an example involving two opinion tuples illustrates the definition of SSA. SSA is more challenging than other related tasks because it requires identifying all four elements of the tuple and may involve overlapping or discontinuous elements. For example, aspect-based sentiment analysis (Pontiki et al., 2014, 2015, 2016; Li et al., 2021b) mostly identifies flat aspect and opinion terms, and opinion mining (Katiyar and Cardie, 2016; Xia et al., 2021) identifies opinion tuples without the sentiment polarity.

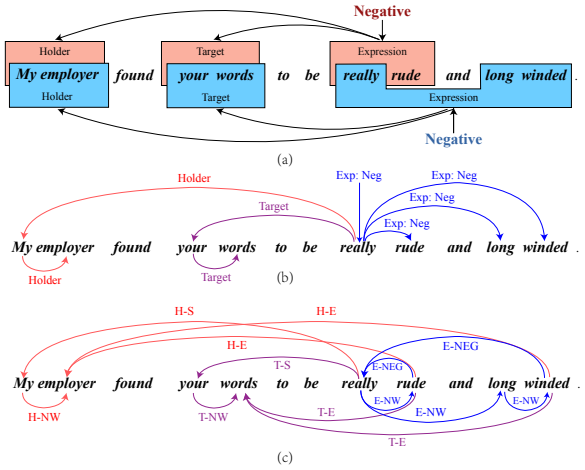

Figure 1: **(a)** An example of structured sentiment analysis, which contains both overlapping and discontinuous situations about two *expressions*. **(b)** Ambiguous *expressions* when using head-first bi-lexical dependency parsing method (Barnes et al., 2021). **(c)** Our proposed unified dependency parsing formulation. E-NEG implies the edge where ending connects to starting (e.g., winded → really) as an *expression* with negative polarity. H-S/E denotes the starting/ending boundary of *holder* and *expression*, and T-S/E denotes the same for *target* and *expression*. ∗-NW indicates the next word.

Most of the existing methods cast the SSA task as the bi-lexical dependency parsing problem. Unfortunately, the conversion is lossy, as it cannot address issues of overlap and discontinuity concurrently. For the example in Figure 1(a), there exist two overlapping[2] e[3], i.e., {*really*, *rude*} and {*really*, *long*, *winded*}, and the latter is discontinuous. Barnes et al. (2021) proposed a dependency parsing method namely head-first as illustrated in Figure 1(b). However, inherent ambiguity occurs in the dependency graph, as the method incorrectly predicts two overlapping e as one single e (i.e.,

*Corresponding author.
[1]Code and datasets are available at https://github.com/zzp-seeker/USSA.
[2]Without loss of generality, "nested" can be considered as a special case of "overlapping" (Fei et al., 2020).
[3]To simplify following explanations, we use h, t, e and p to represent holder, target, expression, and polarity, respectively.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14340–14353
July 9-14, 2023

Dataset	Overlap		Discontinuity	
	#	%	#	%
$\text{NoReC}_{\text{Fine}}$	2178	19.6	1080	9.7
$\text{MultiB}_{\text{EU}}$	0	0	164	7.1
$\text{MultiB}_{\text{CA}}$	3	0.1	113	4.1
MPQA	403	1.4	0	0
DS_{Unis}	18	1.7	102	9.9

Table 1: Count and percentage of tuples in which the entities involve overlapping or discontinuous issues.

{*really*, *rude*, *long*, *winded*}). In other words, the method may not be able to distinguish between two overlapping entities[4] in SSA. Another dependency parsing method proposed by Shi et al. (2022) aims to identify the starting and ending positions of boundaries, but cannot identify discontinuous entities. Statistics on benchmark datasets show the amount of opinion tuples involving overlapping or discontinuous problem in Table 1. Therefore, these two problems cannot be ignored for SSA task and it remains a challenge to design an effective and unified dependency parsing method.

To resolve the kernel challenges (i.e., overlap and discontinuity) existing in SSA, we carefully construct a novel bi-lexical dependency parsing graph as shown in Figure 1(c). The graph comprises two types of edge: ***Relation Prediction*** (RP) and ***Token Extraction*** (TE). RP mainly handles entity boundary identification and relation prediction, and it solves the overlap problem. Specifically, E-POS/NEG/NEU edge connects ending and starting words (e.g., winded $\rightarrow$ really) as an e with sentiment polarity. H-S/E edge marks the starting/ending boundary of h and e, and T-S/E edge is the same for t and e. Another edge type TE identifies all tokens within a given entity boundary, resolving the discontinuity problem. Specifically, $*$-NW edges indicate the next word, meaning that the two words are consecutively joined as a segment of one entity.

Furthermore, we convert our proposed dependency parsing graph to a unified 2D table filling scheme, namely USSA as illustrated in Figure 2. Specifically, we use the start position of each edge as the x-coordinate, the end position as the y-coordinate, and the type of edge as the relationship label in the table. Thus, the table is divided into lower and upper triangular regions, corresponding to RP and TE, respectively.

Based on the USSA scheme, we further develop

[4]In SSA, an entity stands for a holder/target/expression.

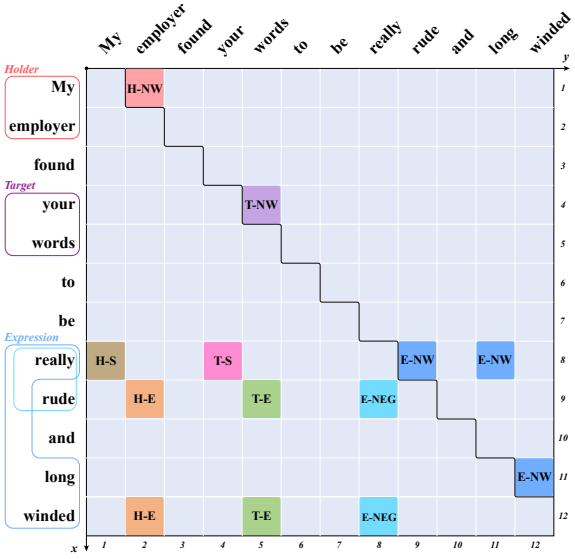

Figure 2: An example to show the conversion from bi-lexical dependency parsing to 2D table filling. The table is divided into lower and upper triangular regions for relation prediction and token extraction, respectively.

a model for SSA. First, multilingual BERT and bi-directional LSTM (BiLSTM) are used to provide contextualized word representations, based on which we construct a 2-Dimensional (2D) table for word pairs. Then, we observe that the relations have a strong correlation in the abscissa and ordinate of the table as shown in Figure 2. we propose a bi-axial attention module to effectively capture these correlations. Finally, a predictor is employed to determine the relations between word pairs.

We conduct extensive experiments on five benchmarks, including $\text{NoReC}_{\text{Fine}}$ (Øvrelid et al., 2020), $\text{MultiB}_{\text{EU}}$, $\text{MultiB}_{\text{CA}}$ (Barnes et al., 2018), MPQA (Wiebe et al., 2005) and DS_{Unis} (Toprak et al., 2010). Our model demonstrates superior performance on all datasets, establishing a new SOTA method for SSA task.

Our contributions are highlighted as follows:

- We propose a bi-lexical dependency parsing graph and convert it to a unified 2D table filling scheme, USSA, which solves the kernel challenge issues of overlap and discontinuity in SSA.
- We present an effective model to well collaborate with USSA scheme, which utilizes proposed bi-axial attention module to better capture the correlations of relations in the table.
- We conduct extensive experiments on five benchmark datasets and the results demonstrate the effectiveness of our model. The source code is released for knowledge sharing.

2 Related Work

Structured Sentiment Analysis (SSA) can be divided into several sub-tasks, including extracting entities, determining the relationship between the entities, and assigning polarity. Some previous research in *Opinion Mining* (OM) has focused on extracting holders, targets, and expressions and identifying their relations, mainly utilizing the MPQA dataset (Esuli et al., 2008). Previous studies have explored different methods to tackle this task, like a BiLSTM-CRF model (Katiyar and Cardie, 2016) that predicts the word-wise opinion role label and identifies the relations, an end-to-end BERT-based model (Quan et al., 2019), a transition-based approach (Zhang et al., 2020a) using pre-defined actions, and a unified span-based model (Xia et al., 2021) that addresses overlap issues. All of these approaches, however, ignore the sentiment polarity classification subtask.

In *Aspect Based Sentiment Analysis* (ABSA), several studies have attempted to unify multiple subtasks. Some examples include *Aspect and Opinion Term Co-Extraction* (AOTE) (Wang et al., 2016, 2017; Dai and Song, 2019; Wang and Pan, 2019; Chen et al., 2020; Wu et al., 2020) which combines target and expression extraction tasks, *Aspect-Sentiment Pair Extraction* (ASPE) (Ma et al., 2018; Li et al., 2019a,b; He et al., 2019) which combines target extraction and sentiment classification, and most recent *Aspect Sentiment Triplet Extraction* (ASTE) (Peng et al., 2020) which further integrates multiple subtasks. These methods can generally be categorized into three groups: Pipeline (Peng et al., 2020; Fan et al., 2019), End-to-End (Zhang et al., 2020b; Xu et al., 2020; Wu et al., 2020; Chen et al., 2021b; Yan et al., 2021; Xu et al., 2021; Chen et al., 2022) and MRC-based (Mao et al., 2021; Chen et al., 2021a; Zhai et al., 2022). However, these methods primarily focus on flat entities and ignore holder extraction.

To this aim, Barnes et al. (2021) originally cast the SSA task as a bi-lexical dependency parsing problem. Nonetheless, as aforementioned in Section 1, the conversion is lossy because it cannot distinguish between two overlapping entities. To address this issue, Shi et al. (2022) proposed another parsing method but unfortunately it cannot identify the discontinuous entities. Samuel et al. (2022) identified the issue of nest and proposed to decode the sentiment graph from the text directly. Therefore, we propose a novel dependency parsing method that can handle overlapping and discontinuous entities simultaneously. We seek to convert the parsing graph to a 2D table filling scheme in order to take advantage of the success of table filling methods (Wang et al., 2020b; Li et al., 2022; Cao et al., 2022) in various fields of NLP.

Type	#	Relation	Meaning of word pair (w_i, w_j)
Relation Prediction	1	E-POS	boundary words of *expression* with positive/negative/neutral polarity
	2	E-NEG	
	3	E-NEU	
	4	H-S	starting or ending boundary of *holder* and corresponding *expression*
	5	H-E	
	6	H-SE	
	7	T-S	starting or ending boundary of *target* and corresponding *expression*
	8	T-E	
	9	T-SE	
Token Extraction	10	E-NW	specific tokens of *expression/holder/target* by indicating w_j is the Next Word of w_i
	11	H-NW	
	12	T-NW	
	13	$\perp$	no above relations

Table 2: The meanings of 13 relations employed in our proposed USSA scheme. *Relation Prediction* and *Token Extraction* are located at lower and upper triangular regions in the table, and solve the overlap and discontinuity problem, respectively.

3 Unified Table Filling Scheme

In this section, we introduce the problem formulation of the SSA task, explain the table filling scheme USSA, and show how to decode opinion tuples from the USSA tagging results.

3.1 Problem Formulation

The objective of SSA is to extract a collection of opinion tuples $\mathcal{T} = \{(h, t, e, p)_m\}_{m=1}^{|\mathcal{T}|}$ from a given input sentence $X = \{w_1, w_2, \cdots, w_N\}$ with N tokens, where h, t, e denote *holder*, *target* and *expression* respectively. The sentiment polarity p of the expression belongs to a sentiment label set, i.e. $\{positive, neutral, negative\}$. The datasets include the challenges posed by discontinuous entities, overlapping counterparts of different tuples, and the presence of null holders and targets.

3.2 Table Filling Scheme

To address the SSA task, USSA uses 13 types of relations between word-pair (w_i, w_j) as shown in Table 2. The table is divided into the lower and upper triangular regions, with the lower region used for relation prediction and the upper region used for token extraction, as depicted in Figure 2.

Relation Prediction (RP) aims to identify the relations between entities and assign the sentiment

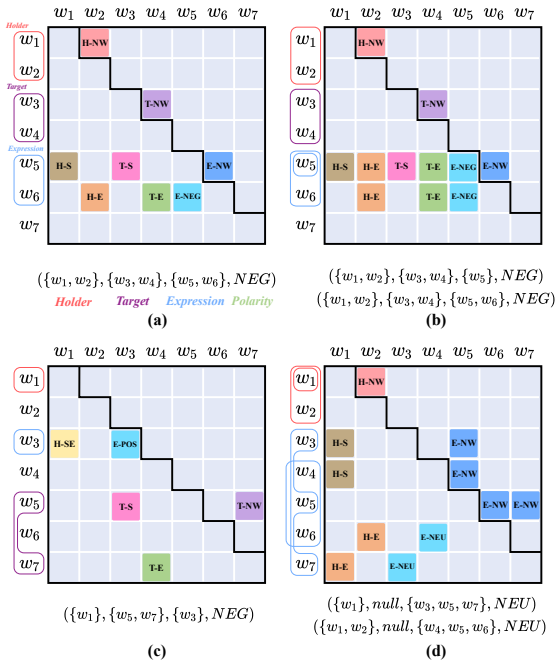

Figure 3: Four decoding cases. **(a)** Flat case. All entities are flat. **(b)** Overlapping case. The expression $\{w_5\}$ is nested in $\{w_5, w_6\}$. **(c)** Discontinuous case. The target $\{w_5, w_7\}$ is discontinuous. **(d)** Complex case. Two expressions $\{w_3, w_5, w_7\}$ and $\{w_4, w_5, w_6\}$ are overlapping, and the former is discontinuous.

polarity. Specifically, ***E-POS/E-NEG/E-NEU*** indicates the starting and ending boundaries of an expression with positive/negative/neutral polarity. ***H-S/H-E/H-SE*** indicates the position of holder and corresponding expression, where *S* and *E* denotes starting and ending positions, respectively. *SE* indicates that the entity consists of only one token, and has the same starting and ending position. In order to ensure that the cell is located in the lower triangle of the table, it is noted that the larger position is set as *x*-coordinate and the smaller is set as *y*-coordinate. ***T-S/T-E/T-SE*** is used in the same manner as the holder for a target.

Token Extraction (TE) aims to extract specific tokens and combine them as an entity based on the entity boundaries obtained from RP. ***E-NW/H-NW/T-NW*** indicates the next word for the expression/holder/target, meaning the pair of words are to be successively joined as a segment of one entity.

3.3 Opinion Tuple Decoding

The overall decoding algorithm is to first identify the boundary words of each holder, target, and expression in the lower triangle region of the table, and then identify the specific tokens in the upper triangle region. First, {*E-POS, E-NEG, E-NEU*} is used to find all boundary words of expression with sentiment polarity. Second, according to {*H-S, H-E, H-SE*} and {*T-S, T-E, T-SE*}, we identify the boundary words of the holders and targets corresponding to the expression, respectively. Finally, we extract the specific tokens of holder, target and expression according to {*E-NW, H-NW, T-NW*} and the corresponding entity boundary. Thus, we collect sentiment tuples (*h,t,e,p*). Figure 3 generally illustrates four decoding cases from easy to difficult.

(a) Flat Case. The boundary words w_5 and w_6 of expression with a negative sentiment polarity can be identified by *E-NEG*. Then according to *H-S* and *H-E*, we can detect the boundary words of holder are w_1 and w_2. Similarly, the boundary words of target are w_3 and w_4. Finally, three paths "$w_1 \rightarrow w_2$", "$w_3 \rightarrow w_4$" and "$w_5 \rightarrow w_6$" are detected as specific words according to the $*-NW$ relations and form a sentiment tuple.

(b) Overlapping case. There are two overlapping expressions and they can be distinguished by two *E-NEG* relations. Therefore, RP relation type contributes to the overlapping issue.

(c) Discontinuous case. There is one discontinuous target in the case. One path "$w_5 \rightarrow w_7$" can be found according to the *T-NW* relation. Therefore, TE relation type can help handling the discontinuous problem.

(d) Complex case. Consider the complex and rare case, where there are two overlapping expressions $\{w_3, w_5, w_7\}$ and $\{w_4, w_5, w_6\}$, and the former is discontinuous. If only use RP, discontinuous expression will incorrectly identified as continuous one (i.e., $\{w_3, w_4, w_5, w_6, w_7\}$). If only use TE, it is impossible to identify correct expressions because we can find four paths in the ambiguous case. Therefore, we can obtain correct tuples by collaboratively using both relation types.

4 Model Structure

This section elaborates upon our model, as depicted in Figure 4, which is designed to effectively integrate the USSA scheme. Our model is mainly composed of four components: the encoder layer, the word-pair representation layer, the refining strategy, and the prediction layer.

4.1 Encoder Layer

Given the input sentence $X = \{w_1, w_2, \cdots, w_N\}$ with N tokens, the encoding layer outputs

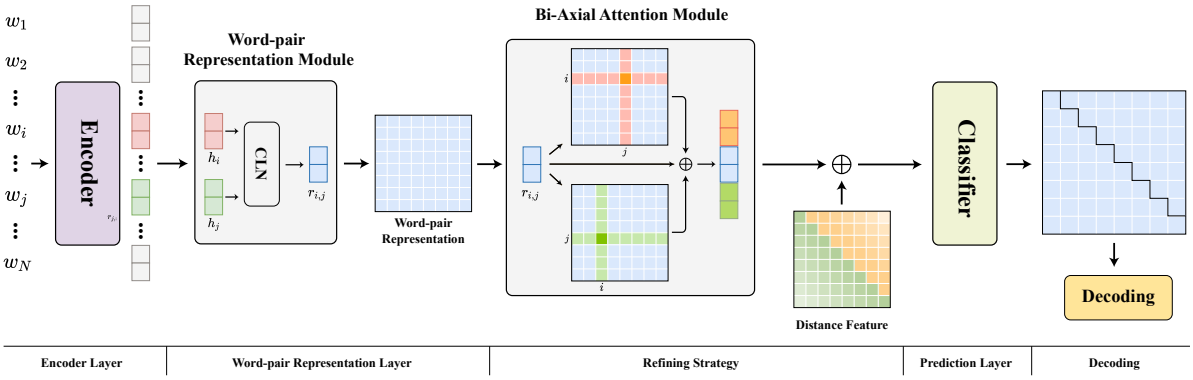

Figure 4: Overall architecture of our end-to-end model. CLN stands for conditional layer normalization.

the hidden representation sequence $H = \{h_1, h_2, \cdots, h_N\}$, leveraging BiLSTM as the sentence encoder. Following previous work, we further enhance the token representations with a pre-trained contextualized embeddings from frozen multilingual BERT (Devlin et al., 2019). Note that we do not use part-of-speech (POS), lemma, and character-level embedding, which may put our model at a disadvantage in comparison to other models that do.

4.2 Word-pair Representation Layer

Evidently, the relation of USSA is asymmetric ($(w_i, w_j) \neq (w_j, w_i)$). Inspired by Yu et al. (2021) and Wang et al. (2021), we utilize Conditional Layer Normalization (CLN) to model the conditional word-pair representation $\mathbf{R}$ as,

$$\begin{aligned} r_{i,j} &= \mathrm{CLN}(h_i, h_j) \\ &= \gamma_i \odot \left(\frac{h_j - \mu}{\sigma} \right) + \lambda_i \end{aligned} \tag{1}$$

where $\odot$ denotes the element-wise product. In addition, scale factor γ_i and shift factor λ_i can incorporate extra contextual information, and two parameters μ and σ are the mean and standard deviation of h_j, i.e.,

$$\gamma_i = \mathrm{W}_\gamma h_i + b_\gamma, \quad \lambda_i = \mathrm{W}_\lambda h_i + b_\lambda \tag{2}$$

and

$$\mu = \frac{1}{d} \sum_{k=1}^{d} h_{jk}, \quad \sigma = \sqrt{\frac{1}{d} \sum_{k=1}^{d} (h_{jk} - \mu)^2} \tag{3}$$

where h_{jk} denotes the k-th dimension of h_j. W_γ, W_λ, b_γ and b_λ are learnable parameters.

4.3 Refining Strategy

Bi-Axial Attention Module. The relations in USSA exhibit a strong correlation in the rows and columns of the table. As an example, Figure 2 illustrates that for the *E-NEG* at position (12,8), the corresponding *H-S*/*H-E*/*T-S*/*T-E* must be located in row 8 or row 12 or column 8 or column 12 if it exists, and there must be *E-NW* relations in row 8 and in column 12. We propose to adopt bi-axial attention module to capture the correlation of relations and ensure the global connection, drawing inspiration from the success of axial attention (Ho et al., 2019; Wang et al., 2020a; Huang et al., 2019) in computer vision. First, we define a single axial attention as follows,

$$\begin{aligned} a_{i,j} = &\ \mathrm{MultiHead}(r_{i,j}, row_i, row_i) + \\ &\ \mathrm{MultiHead}(r_{i,j}, col_j, col_j) \end{aligned} \tag{4}$$

where $\mathrm{MultiHead}$, row_i, col_j represent multi-head attention, the i-th row, and the j-th column of word-pair representation $\mathbf{R}$, respectively. Then we utilize another symmetric axial attention and the word-pair representation itself to construct the contextual representation $\mathbf{C}$ as,

$$c_{i,j} = a_{i,j} \oplus r_{i,j} \oplus a_{j,i} \tag{5}$$

where $\oplus$ denotes the concatenation operation.

Feature Enhancement. To further improve the representation, we introduce the distance feature as shown in Figure 4. In light of the fact that the relation in USSA is sensitive to the relative distance of word pairs (e.g., the greater the NW span, the more words are spaced for the next word), we use distance feature to represent the relative distance information. Additionally, it helps to distinguish between the lower and upper triangular regions.

Dataset	Split	Sentences	Holders			Targets			Expressions			POS	NEU	NEG
			all	over.	dis.	all	over.	dis.	all	over.	dis.			
NoReC$_{\text{Fine}}$	train	8634	898	0	0	6778	0	39	8448	1655	781	5684	0	2756
	dev	1531	120	0	0	1152	0	5	1432	261	131	988	0	443
	test	1272	110	0	0	993	0	6	1235	262	125	875	0	358
MultiB$_{\text{EU}}$	train	1064	205	0	4	1285	0	23	1684	0	91	1406	0	278
	dev	152	33	0	0	153	0	1	204	0	15	168	0	36
	test	305	58	0	6	337	0	4	440	0	23	375	0	65
MultiB$_{\text{CA}}$	train	1174	169	0	1	1695	0	23	1981	0	61	1272	0	708
	dev	168	15	0	0	211	0	1	258	3	6	151	0	107
	test	336	52	0	0	430	0	8	518	0	18	313	0	204
MPQA	train	5873	1431	0	0	1487	241	0	1715	6	0	671	337	698
	dev	2063	414	0	0	503	80	0	581	2	0	223	126	216
	test	2112	434	0	0	462	80	0	518	0	0	159	82	223
DS$_{\text{Unis}}$	train	2253	65	0	0	836	16	0	836	0	82	349	104	383
	dev	232	9	0	0	104	0	0	104	0	8	31	16	57
	test	318	12	0	0	142	2	0	142	0	12	59	12	71

Table 3: Statistics of the datasets, including the number of sentences per split, as well as the number of holder, target, and expression annotations. Additionally, we include the number of overlapping (over.) and discontinuous (dis.) entities, as well as the distribution of polarity in each dataset.

Then the final representation $\mathbf{V}$ is obtained as,

$$v_{i,j} = c_{i,j} \oplus d_{i-j} \tag{6}$$

where d_{i-j} is the relative distance embedding.

4.4 Prediction Layer

To obtain the label probability distribution $p_{i,j}$ for each cell in the table, we feed the refined word pair representation $v_{i,j}$ into a feed-forward network (FFN) and input $h_{i,j}$ into the biaffine predictor as an enhancement.

FFN Predictor. For the word pair representation $v_{i,j}$, we utilize an FFN to obtain the relation score as,

$$s^f_{i,j} = \text{FFN}_f(v_{i,j}) \tag{7}$$

where $s^f_{i,j} \in \text{R}^m$ is the relation score, and m is the number of relation type.

Biaffine Predictor. Biaffine has proven effective for dependency parsing (Dozat and Manning, 2017), and it can work collaboratively with FFN predictor for relation classification according to previous research (Li et al., 2021a, 2022). We use biaffine module in our model to obtain the relation score $s^b_{i,j}$ between the word pair (w_i, w_j) as an enhancement, i.e.,

$$h^a_i = \text{FNN}_a(h_i) \tag{8}$$

$$h^b_j = \text{FNN}_b(h_j) \tag{9}$$

$$s^b_{i,j} = {h^a_i}^{\text{T}} U_1 h^b_j + U_2 \left(h^a_i \oplus h^b_j\right) + b \tag{10}$$

where U_1, U_2 and b are trainable weights and bias. Thus, the relation score $s^b_{i,j} \in \text{R}^m$ is obtained. Finally, the label probability distribution is calculated by combining the relation scores $s^f_{i,j}$ and $s^b_{i,j}$ as,

$$p_{i,j} = \text{softmax}\left(\alpha s^f_{i,j} + (1-\alpha) s^b_{i,j}\right) \tag{11}$$

where α is hyper-parameter used to adjust the influence of the corresponding predictor.

4.5 Loss Function

Our objective is to minimize the following cross-entropy loss as follows,

$$\mathcal{L} = -\sum_{i}^{N}\sum_{j}^{N}\sum_{r\in\mathcal{R}} \mathbb{I}(y_{ij} = r)\log(p_{i,j|r}) \tag{12}$$

where N is the number of tokens in the sentence and $\mathcal{R}$ is pre-defined relation set in USSA.

5 Experiments

5.1 Datasets and Configuration

Following the previous work, we conduct experiments on five benchmark datasets in four languages. The statistics are shown in Table 3. NoReC$_{\text{Fine}}$ (Øvrelid et al., 2020) is a professional reviews dataset in Norwegian. MultiB$_{\text{EU}}$ and MultiB$_{\text{CA}}$ (Barnes et al., 2018) annotates hotel views in Basque and Catalan, respectively. MPQA (Wiebe et al., 2005) contains English news wire text and the content of DS$_{\text{Unis}}$ (Toprak et al., 2010) is online universities reviews in English.

Dataset	Model	Span			Sent. Graph	
		Holder F1 ↑	Target F1 ↑	Exp. F1 ↑	NSF1 ↑	SF1 ↑
$\text{NoReC}_{\text{Fine}}$	RACL-BERT (Chen and Qian, 2020)	_	47.2	56.3	_	_
	Head-first (Barnes et al., 2021)	51.1	50.1	54.4	37.0	29.5
	Head-final (Barnes et al., 2021)	60.4	**54.8**	55.5	39.2	31.2
	Frozen PERIN (Samuel et al., 2022)	48.3	51.9	57.9	41.8	35.7
	TGLS (Shi et al., 2022)	60.9	53.2	61.0	46.4	37.6
	USSA (Ours)	**66.3**	54.3	**61.4**	**47.7**	**39.6**
$\text{MultiB}_{\text{EU}}$	RACL-BERT (Chen and Qian, 2020)	_	59.9	72.6	_	_
	Head-first (Barnes et al., 2021)	60.4	64.2	73.9	58.0	54.7
	Head-final (Barnes et al., 2021)	60.5	64.0	72.1	58.2	54.7
	Frozen PERIN (Samuel et al., 2022)	55.5	58.5	68.8	53.1	51.3
	TGLS (Shi et al., 2022)	62.8	65.6	75.2	61.1	58.9
	USSA (Ours)	**63.4**	**66.9**	**75.4**	**63.5**	**60.4**
$\text{MultiB}_{\text{CA}}$	RACL-BERT (Chen and Qian, 2020)	_	67.5	70.3	_	_
	Head-first (Barnes et al., 2021)	43.0	72.5	71.1	62.0	56.8
	Head-final (Barnes et al., 2021)	37.1	71.2	67.1	59.7	53.7
	Frozen PERIN (Samuel et al., 2022)	39.8	69.2	66.3	60.2	57.6
	TGLS (Shi et al., 2022)	47.4	73.8	71.8	64.2	59.8
	USSA (Ours)	**47.5**	**74.2**	**72.2**	**67.4**	**61.0**
MPQA	RACL-BERT (Chen and Qian, 2020)	_	20.0	31.2	_	_
	Head-first (Barnes et al., 2021)	43.8	51.0	**48.1**	24.5	17.4
	Head-final (Barnes et al., 2021)	46.3	49.5	46.0	26.1	18.8
	Frozen PERIN (Samuel et al., 2022)	44.0	49.0	46.6	30.7	23.1
	TGLS (Shi et al., 2022)	44.1	51.7	47.8	28.2	21.6
	USSA (Ours)	**47.3**	**58.9**	48.0	**36.8**	**30.5**
DS_{Unis}	RACL-BERT (Chen and Qian, 2020)	_	44.6	38.2	_	_
	Head-first (Barnes et al., 2021)	28.0	39.9	40.3	31.0	25.0
	Head-final (Barnes et al., 2021)	37.4	42.1	45.5	34.3	26.5
	Frozen PERIN (Samuel et al., 2022)	13.8	37.3	33.2	24.5	21.3
	TGLS (Shi et al., 2022)	43.7	49.0	42.6	36.1	31.1
	USSA (Ours)	**44.2**	**50.2**	**46.6**	**38.0**	**33.2**

Table 4: Experiment results on five benchmark datasets for SSA task.

We obtain the frozen token representations from the pre-trained BERT-multilingual-base to ensure a fair comparison with other methods. Furthermore, we use 4-layer BiLSTMs with an output size of 768. We train our model for 60 epochs with a linear warm-up for 10% of the training steps and save the model parameters based on the highest SF1 score on the development set. We use an NVIDIA A100 to train the model for an average of 45 minutes. The reported results are the averages from five runs with different random seeds. See Appendix A for more details.

5.2 Baseline Methods

We compare our proposed method with five state-of-the-art baselines. **RACL-BERT** (Chen and Qian, 2020) is a relation aware collaborative learning framework which allows the subtasks of ABSA to work coordinately. Barnes et al. (2021) employ it as a baseline for SSA. **Head-first** and **Head-final** (Barnes et al., 2021) are two different bi-lexical dependency parsing methods that use a re-implementation of the neural parser (Dozat and Manning, 2018). **Frozen PERIN** (Samuel et al., 2022) applies PERIN (Samuel and Straka, 2020), a graph-based parser to model a superset of graph features into a frozen XLM-R (Conneau et al., 2020) backbone. **TGLS** (Shi et al., 2022) is a bi-lexical dependency parsing method and it is equipped with the graph attention network.

5.3 Evaluation Metrics

Following the previous work (Samuel et al., 2022), we mainly use **Sentiment Graph F1 (SF1)** to evaluate our models. SF1 defines a sentiment tuple as a true positive when it is an exact match at graph-

	NoReC$_{\text{Fine}}$	MultiB$_{\text{EU}}$	MultiB$_{\text{CA}}$	MPQA	DS$_{\text{Unis}}$
Full Model	**39.58**	**60.39**	**61.02**	**30.46**	**33.19**
w/o bi-axial att.	38.67 (-0.91)	59.61 (-0.78)	60.03 (-0.99)	29.83 (-0.63)	32.17 (-1.02)
w/ single-axial att.	39.32 (-0.26)	59.91 (-0.48)	60.44 (-0.58)	29.98 (-0.48)	32.88 (-0.31)
w/o distance	39.41 (-0.17)	60.01 (-0.38)	60.67 (-0.35)	30.21 (-0.25)	32.99 (-0.20)
w/o biaffine	39.22 (-0.36)	60.11 (-0.28)	60.84 (-0.18)	29.99 (-0.47)	33.02 (-0.17)
w/o FFN	37.72 (-1.86)	59.23 (-1.16)	59.60 (-1.42)	29.36 (-1.10)	31.59 (-1.60)
w/o $*$-NW	38.29 (-1.29)	59.82 (-0.57)	60.41 (-0.61)	30.42 (-0.04)	32.18 (-1.01)

Table 5: The SF1 scores of ablation study.

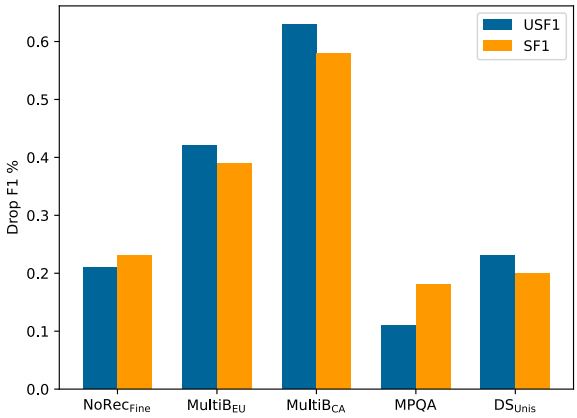

Figure 5: The decline in performance when replacing the bi-axial attention module with CNN.

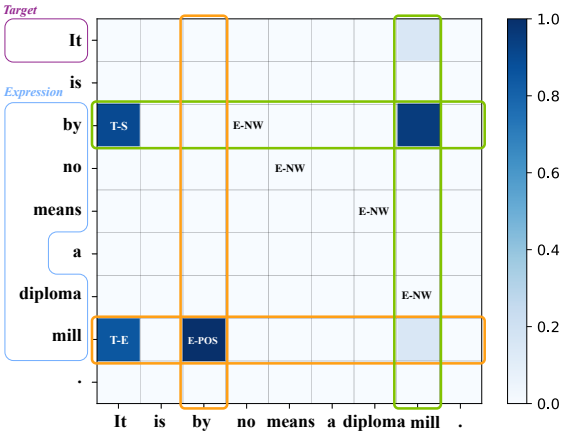

Figure 6: Visualization of the bi-axial attention scores applied to the E-POS cell.

level, weighting the overlap between the predicted and gold spans for each span, and averaging across all three spans. We also include **Holder F1**, **Target F1**, and **Exp. F1** for token extraction of *Holders*, *Targets*, and *Expressions*, as well as **Nonpolarity Sentiment Graph F1 (NSF1)** for further analysis.

5.4 Main Results

In Table 4, we compare our USSA with other baselines using multiple metrics. We find that our USSA generally outperforms the other baselines in terms of the Span F1 metric across all datasets, and it surpasses the performance of suboptimal method by an average of 1.47% F1 score. It includes significant improvements, such as 7.2% F1 score for extracting targets on MPQA and 5.4% F1 score for extracting holders on NoReC$_{\text{Fine}}$. However, the performance of our USSA in extracting targets is slightly weaker, with a 0.5% lower F1 score. Considering the Sentiment Graph metric, which is crucial for comprehensively evaluating entity, relation and polarity predictions, our USSA consistently outperforms all other methods in both NSF1 and SF1. Compared with another strong baseline TGLS, our USSA surpasses its performance by averages 3.48 NSF1 score and 3.14% SF1 score, despite of not using POS, lemma, or character-level embedding. The improvement is attributed to our USSA's ability to effectively address the issues of overlap and discontinuity simultaneously.

6 Discussion

In this section, we will conduct a deeper analysis to answer the following questions.

6.1 Are the components of the model valid?

Table 5 presents the findings of ablation experiments. The results reveal that the bi-axial attention module is useful for good performance, as its removal resulted in an obvious decline in performance across all five datasets. On the other hand, substituting the bi-axial attention module with a single one or removing the distance feature has a less pronounced effect on performance. Furthermore, we find that while the FFN predictor played a dominant role, the biaffine predictor also makes a positive impact, with gains up to 0.47% observed at most. Lastly, discarding the $*$-NW relations cause a noticeable drop in F1 scores across all datasets, particularly on NoReC$_{\text{Fine}}$ ($\downarrow$1.29%) and DS$_{\text{Unis}}$ ($\downarrow$1.01%). This is because these datasets contain a higher proportion of discontinuous entities, and without the $*$-NW relations, such entities would be incorrectly identified as continuous ones. In short, the results demonstrate the effectiveness of each module and emphasize the significance of the $*$-NW relations.

6.2 Is bi-axial attention module effective?

Previous research has demonstrated the effectiveness of convolutional neural networks (CNNs) in table filling methods (Li et al., 2022; Yan et al., 2022). However, when the table is large, it can be challenging for CNNs to fast capture global information (Peng et al., 2021). We conduct a direct comparison with the CNN method used in (Li et al., 2022) as shown in Figure 5. The results indicate that the performance of CNNs decreases across all five datasets, and it is likely due to the fact that many sentences in SSA tasks are long. In addition, we visualize the bi-axial attention scores applied to the E-POS cell as shown in Figure 6. The visualization shows the attention on related relations of E-POS, such as T-S and T-E. To sum up, bi-axial attention mechanism can effectively help identify relations in the table.

7 Conclusion

In this paper, we propose a novel bi-lexical dependency parsing graph and convert it to a unified 2D table-filling scheme, namely USSA to address the overlapping and discontinuous issues simultaneously. We also develop a model that includes a novel bi-axial attention module to better refine the word-pair representation. Additionally, our proposed framework may serve as an inspiration for other tasks involving the extraction of tuples that both present overlap and discontinuity challenges.

Limitations

Our approach has proven to be superior to previous methods on multiple public benchmark datasets. However, one major disadvantage of the table filling method is the increased training time and memory usage. The computational resources are required for the 2D table representation of word-pair relations for constructing and storing the table. In comparison, using a sequence representation as input could be generally more efficient. Our approach also faces the computational challenge.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62076032. We appreciate constructive feedback from the anonymous reviewers for improving the final version of this paper.

References

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. 2022. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524, Online. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021a. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12666–12674.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021b. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1474–1483, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5268–5277, Florence, Italy. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Andrea Esuli, Fabrizio Sebastiani, and Ilaria Urciuoli. 2008. Annotating expressions of opinion and emotion in the Italian content annotation bank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zhifang Fan, Zhen Wu, Xin-Yu 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fei, Yafeng Ren, and Donghong Ji. 2020. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311.

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, Florence, Italy. Association for Computational Linguistics.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2019. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 603–612. IEEE.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10965–10973.

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021a. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, Online. Association for Computational Linguistics.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021b. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13543–13551.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. 2021. Conformer: Local features coupling global representations for visual recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 357–366. IEEE.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris 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, Denver, Colorado. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Wei Quan, Jinli Zhang, and Xiaohua Tony Hu. 2019. End-to-end joint opinion role labeling with bert. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2438–2446.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–478, Dublin, Ireland. Association for Computational Linguistics.

David Samuel and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, Dublin, Ireland. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. 2020a. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 108–126. Springer.

Wenya Wang and Sinno Jialin Pan. 2019. Transferable interactive memory network for domain adaptation in fine-grained opinion extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7192–7199.

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, Austin, Texas. Association for Computational Linguistics.

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 Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322. AAAI Press.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020b. TPLinker: Single-stage joint extraction of entities

and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous named entity recognition as maximal clique discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 764–774, Online. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 39(2-3):165–210.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A unified span-based approach for opinion mining with syntactic constituents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1795–1804, Online. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Hang Yan, Yu Sun, Xiaonan Li, and Xipeng Qiu. 2022. An embarrassingly easy but strong baseline for nested named entity recognition. *CoRR*, abs/2208.04534.

Bowen Yu, Zhenyu Zhang, Jiawei Sheng, Tingwen Liu, Yubin Wang, Yucheng Wang, and Bin Wang. 2021. Semi-open information extraction. In *Proceedings of the Web Conference 2021*, WWW '21, page 1661–1672, New York, NY, USA. Association for Computing Machinery.

Zepeng Zhai, Hao Chen, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. COM-MRC: A COntext-masked machine reading comprehension framework for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3230–3241, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bo Zhang, Yue Zhang, Rui Wang, Zhenghua Li, and Min Zhang. 2020a. Syntax-aware opinion role labeling with dependency graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3249–3258, Online. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020b. A multi-task learning framework for opinion triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 819–828, Online. Association for Computational Linguistics.

A Hyper-parameter Settings

Global Hyper-parameter Settings

Hyperparameter	Assignment
Contextualized Embedding	mBERT
Embeddings Trainable	False
Num of Epochs	60
Batch Size	16
Hidden LSTM	768
Dim Distance Feature	100
Gradient Accumulation Step	2

Local Hyper-parameter Settings

Dataset	MaxTokenLen	LearningRate	α
$\text{NoReC}_{\text{Fine}}$	150	2e-3	0.650
$\text{MultiB}_{\text{EU}}$	150	2e-3	0.500
$\text{MultiB}_{\text{CA}}$	386	1e-3	0.650
MPQA	210	2e-3	0.725
DS_{Unis}	386	1e-3	0.650

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section on Page 9

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract section and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5.1

☑ B1. Did you cite the creators of artifacts you used?
Section 5.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.1

C ☑ Did you run computational experiments?

Section 5.4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 and Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.1

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.