Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, Donghong Ji[†]
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{shiwenxuan,lifei_csnlp,theodorelee,hao.fei,dhji}@whu.edu.cn

Abstract

The state-of-the-art model for structured sentiment analysis casts the task as a dependency parsing problem, which has some limitations: (1) The label proportions for span prediction and span relation prediction are imbalanced. (2) The span lengths of sentiment tuple components may be very large in this task, which will further exacerbates the imbalance problem. (3) Two nodes in a dependency graph cannot have multiple arcs, therefore some overlapped sentiment tuples cannot be recognized. In this work, we propose nichetargeting solutions for these issues. First, we introduce a novel labeling strategy, which contains two sets of token pair labels, namely essential label set and whole label set. The essential label set consists of the basic labels for this task, which are relatively balanced and applied in the prediction layer. The whole label set includes rich labels to help our model capture various token relations, which are applied in the hidden layer to softly influence our model. Moreover, we also propose an effective model to well collaborate with our labeling strategy, which is equipped with the graph attention networks to iteratively refine token representations, and the adaptive multi-label classifier to dynamically predict multiple relations between token pairs. We perform extensive experiments on 5 benchmark datasets in four languages. Experimental results show that our model outperforms previous SOTA models by a large margin.[1]

1 Introduction

Structured Sentiment Analysis (SSA), which aims to predict a structured sentiment graph as shown in Figure 1(a), can be formulated into the problem of tuple extraction, where a tuple (h, e, t, p) denotes a holder h who expressed an expression e towards a target t with a polarity p. SSA is a more challenging task, because other related tasks only focus on extracting part of tuple components or the text spans of the components are short. For example, *Opinion Role Labeling* (Katiyar and Cardie, 2016; Xia et al., 2021) does not include the extraction of sentiment polarities, and *Aspect-Based Sentiment Analysis* (ABSA) (Pontiki et al., 2014; Wang et al., 2016) extracts the aspect and opinion terms typically consisting of one or two words. The state-of-the-art SSA model is proposed by Barnes et al. (2021), which casts the SSA task as the dependency parsing problem and predicts all tuple components as a dependency graph (Figure 1(b)).

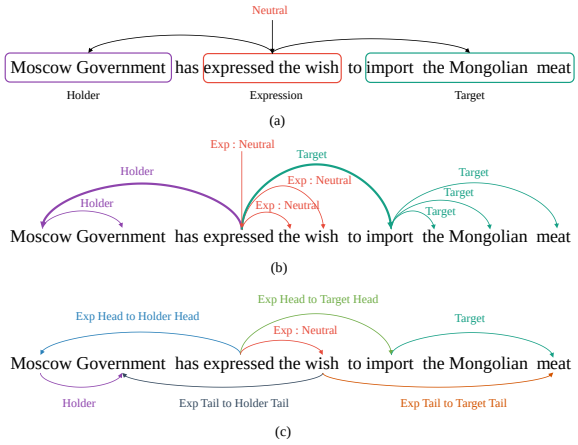

Figure 1: **(a)** An example of structured sentiment analysis. **(b)** The head-first parsing graph proposed by Barnes et al. (2021), where the arcs related to holder(target)-expression linking relations are bold. **(c)** Our proposed *essential label set*, which has more balanced label distribution for holder, target or expression span prediction and their linking relation prediction.

However, their method exists some shortages. Taking Figure 1(b) as example, only 2 arcs (e.g., expressed→import and expressed→Moscow) are related to span linking relation prediction (i.e., the relations between expressions and holders or targets), while much more other arcs are related to span prediction (e.g., import→the and import→meat). Such imbalanced labeling strat-

[†]Corresponding author
[1]Our code is available at https://github.com/Xgswlg/TGLS

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4232 - 4241

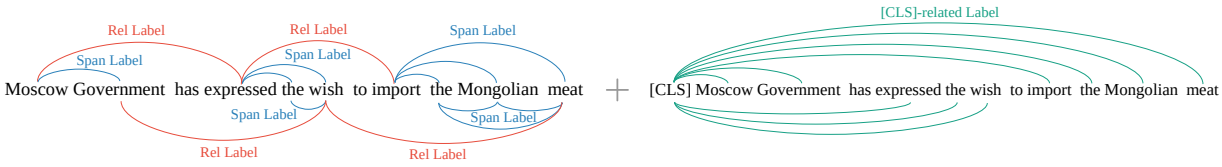

Figure 2: The whole label set contains the labels for span prediction and span relation prediction, as well as the [CLS]-related labels that connect a sentinel [CLS] token with the holder, target and expression tokens.

Dataset	Span Length ≥ 4			Multi Label
	Hoder	Target	Exp.	
NoReC$_{\text{Fine}}$	1.1%	19.2%	56.8%	14.0%
MultiB$_{\text{CA}}$	2.6%	18.4%	21.4%	8.7%
MultiB$_{\text{EU}}$	1.1%	2.7%	15.3%	3.6%
MPQA	19.9%	51.1%	14.5%	1.0%
DS$_{\text{Unis}}$	1.3%	0.8%	13.7%	1.9%

Table 1: Statistics of the proportion of each sentiment component whose span length (in tokens) is greater than or equal to 4, and the proportion of sentences requiring multi-label classification for SSA.

egy will make the model pay more attention on span prediction but less on span relation prediction. Furthermore, since the span lengths of sentiment tuple components may be very large in the SSA task, the label imbalanced problem will become more severe. Besides, the dependency parsing graph is not able to deal with multi-label classification, since it does not allow multiple arcs to share the same head and dependent tokens. Therefore, some overlapped sentiment tuples cannot be recognized. The statistics of span length and multi-label problems are listed in Table 1.

To alleviate the label imbalance problem in Barnes et al. (2021), we propose a novel labeling strategy that consists of two parts: First, we design a set of labels called **essential label set** (Figure 1(c)), which can be considered as the basic label set for decoding SSA tuples, since it only includes the labels to tag the boundary tokens of spans. As seen, the proportion of span prediction labels and span relation prediction labels are relatively balanced, so that we can mitigate the label imbalance problem and meanwhile keep the basic ability of extracting sentiment tuples if the essential label set is learnt in the final prediction layer of our model.

However, the labels related to recognize non-boundary tokens of SSA components are also important. For instance, they can encode the relations between the tokens inside the spans, which may benefit the extraction of the holders, expressions or targets with long text spans. To this end, we design another label set called **whole label set** (Figure 2), which includes richer labels to fully utilize various information such as the relations among boundary tokens, non-boundary tokens, the tokens within a span, the tokens across different spans. Moreover, since the dependency-based method (Barnes et al., 2021) only considers the local relation between each pair of tokens, we add the labels between [CLS] and other tokens related to sentiment tuples into our whole label set, in order to utilize sentence-level global information. Considering that if the whole label set is directly applied on the output label for training, the label imbalance problem may occur again. We instead employ the whole label set in a soft and implicit fashion by applying it on the hidden layer of our model.

To well collaborate with our labeling strategy, we also propose an effective token graph model, namely **TGLS** (**T**oken **G**raph with a novel **L**abeling **S**trategy), which uses rich features such as word, part-of-speech tags and characters as inputs and yields contextualized word representations by BiLSTM and multilingual BERT(Devlin et al., 2018). In the hidden layer, we build a multi-view token graph, which has four views corresponding to different relations in the whole label set and each view is a graph attention network (Veličković et al., 2017) with token representations as the nodes. In the prediction layer, we introduce a novel adaptive multi-label classifier to extract all the sentiment tuples no matter that they are overlapped or not.

We conduct extensive experiments on five benchmarks, including NoReC$_{\text{Fine}}$ (Øvrelid et al., 2020), MultiB$_{\text{EU}}$, MultiB$_{\text{CA}}$ (Barnes et al., 2018), MPQA (Wiebe et al., 2005) and DS$_{\text{Unis}}$ (Toprak et al., 2010). The resluts show that our TGLS model outperforms the SOTA model by a large margin. In summary, our main contributions include:

- We design a novel labeling strategy to address the label imbalance issue in prior work. Concretely, we employ the whole label set and essential label set in the hidden and prediction layer respectively, achieving a balance

between the label variety and label imbalance.

- We propose an effective token graph model to well collaborate with our labeling strategy, which learns the token-token relations via multi-view token graph networks and reasons the labels between each pair of words using the adaptive multi-label classifier for both overlapped and non-overlapped tuple extraction.
- The experimental results show that our model has achieved the SOTA performance in 5 datasets for structured sentiment analysis, especially in terms of the end-to-end sentiment tuple extraction.

2 Related Work

The task of the Structured Sentiment Analysis (SSA) can be divided into sub-tasks such as span extraction of the holder, target and expression, relation prediction between these elements and assigning polarity. Some existing works in *Opinion Mining* used pipeline methods to first extract spans and then the relations mostly on the MPQA dataset (Wiebe et al., 2005). For example, Katiyar and Cardie (2016) propose a BiLSTM-CRF model which is the first such attempt using a deep learning approach, Zhang et al. (2019) propose a transition-based model which identifies opinion elements by the human-designed transition actions, and Xia et al. (2021) propose a unified span-based model to jointly extract the span and relations. However, all of these works ignore the polarity classification sub-task.

In *End2End Aspect-Based Sentiment Analysis* (ABSA), there are also some attempts to unify several sub-tasks. For instance, Wang et al. (2016) augment the ABSA datasets with sentiment expressions, He et al. (2019) make use of this data and models the joint relations between several sub-tasks to learn common features, and (Chen and Qian, 2020) also exploit interactive information from each pair of sub-tasks (target extraction, expression extraction, sentiment classification). However, Wang et al. (2016) only annotate sentiment-bearing words not phrases and do not specify the relationship between target and expression, it therefore may not be adequate for full structured sentiment analysis.

Thus, Barnes et al. (2021) propose a unified approach in which they formulate the structured sentiment analysis task into a dependency graph parsing task and jointly predicts all components of a sentiment graph. However, as aforementioned, this direct transformation may be problematic as it may introduce label imbalance in span and relation prediction. Thus, we propose an effective graph model with a novel labeling strategy in which we employ a whole label set in the hidden layer to softly affect our model, and an essential label set in the prediction layer to address the imbalance issue.

The design of our essential label set is inspired by the Handshaking Tagging Scheme (Wang et al., 2020), which is a token pair tagging scheme for entity and relation extraction. The handshaking tagging scheme involves only the labels related to the boundary tokens and enables a one-stage joint extraction of spans and relations. In our work, we modify the handshaking tagging scheme to use it for SSA. Furthermore, since the component span of this task is relatively long, only utilizing the boundary tokens cannot make full use of the annotation information, so we propose a new label set called whole label set, which together with essential label set constitutes our labeling strategy.

3 Token-Pair Labeling Strategy

3.1 Essential Label Set

Our essential label set only involves the labels related to the boundary tokens, therefore the label proportions for span prediction and span relation prediction are relatively balanced. Given a sentence "*Moscow government has expressed the wish to import the Mongolian meat.*", the essential label set consists of the following labels:

- *Holder*: *Moscow* → *government*
- *Exp:Neutral*: *expressed* → *Moscow*
- *Target*: *import* → *meat*
- *Exp Head to Holder Head*: *expressed* → *Moscow*
- *Exp Tail to Holder Tail*: *wish* → *government*
- *Exp Head to Target Head*: *expressed* → *import*
- *Exp Tail to Target Tail*: *wish* → *meat*

where the *Holder*, *Exp.* and *Target* represent the three components of a sentiment tuple, the *Head* or *Tail* means the start or end token of a component, and the *Neutral* denotes the polarity.

3.2 Whole Label Set

Our whole label set involves both the labels related to boundary and non-boundary tokens, as well as

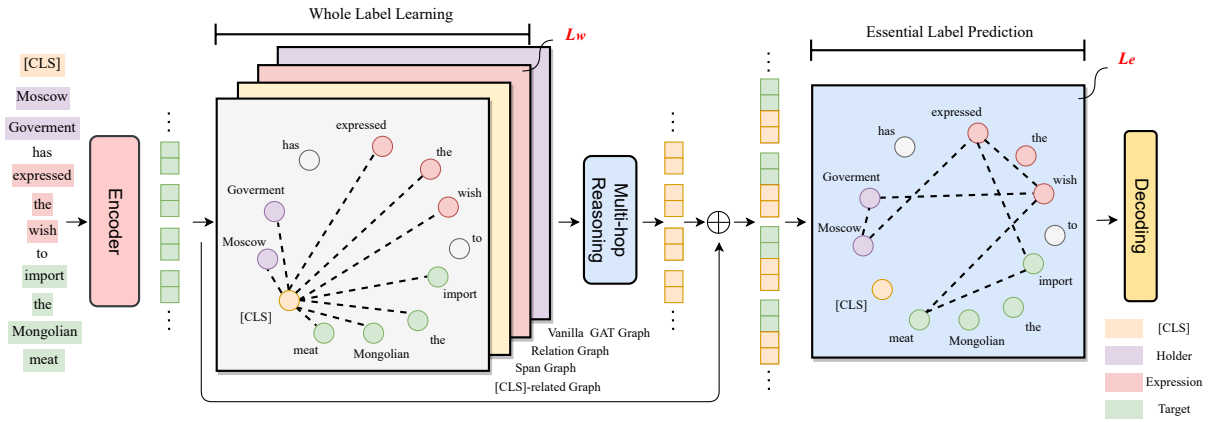

Figure 3: Overall architecture of the our framework. From left to right, the first is an encoder to yield contextualized word representations from input sentences, and the next is a graph layer where we produce attention scoring matrices by whole label prediction. Then we build a multi-hop reasoning layer and refine token representations. Finally, a prediction layer is leveraged for reasoning the relations in essential labels and based on which we decode all components of an opinion tuple.

the labels related to `[CLS]` and all tokens in the sentiment tuples. Thus, our whole label set can be divided into three groups, span labels, relation labels and `[CLS]`-related labels. Given the sentence in Figure 2, the whole label set include the following labels:

- *Span Label*: e.g. *import* → *Mongolian*
- *Rel Label*: e.g. *Moscow* → *expressed*
- *`[CLS]`-related Label*: e.g. *`[CLS]`* → *expressed*

where the span and relation labels make our model be aware of the token relations inside and across the spans of sentiment components, and `[CLS]`-related labels can help our model to capture the sentence-level global information. We apply whole labels in the hidden layer to softly embed the above information into our model, in order to avoid the potential label imbalance issue.

3.3 Decoding

We first decode all the expression-holder and expression-target pairs that meet the constraints of essential label set. In detail, we can get all component spans based on span prediction labels (e.g. *Holder*, *Exp:Neutral* and *Target* labels), then we decode all expression to holder or target pairs as long as it meets one of the corresponding relation prediction labels (e.g. for expression to holder pairs, the labels are *Exp Head to Holder Head* and *Exp Tail to Hoder Tail*). After decoding all the component pairs, we enumerate all possible triples from pairs with the same expression, thus finally decode all the sentiment tuples.

4 Methodology

In this section, We formally present our proposed TGLS model in detail (Figure 3), which mainly consists of four parts, the encoder layer, the multi-view token graph as the hidden layer, the adaptive multi-label classifier as the prediction layer and the hierarchical learning strategy to train our model.

4.1 Encoder Layer

Consider the i^{th} token in a sentence with n tokens, we represent it by concatenating its token embedding $\mathbf{e}_i^{word}$, part-of-speech (POS) embedding $\mathbf{e}_i^{pos}$, lemma embedding $\mathbf{e}_i^{lemma}$, and character-level embedding $\mathbf{e}_i^{char}$ together:

$$\boldsymbol{w}_i = \mathbf{e}_i^{word} \oplus \mathbf{e}_i^{pos} \oplus \mathbf{e}_i^{lemma} \oplus \mathbf{e}_i^{char} \quad (1)$$

where $\oplus$ denotes the concatenation operation. The character-level embedding is generated by the convolution neural networks (CNN) (Kalchbrenner et al., 2014). Then, we employ bi-directional LSTM (BiLSTM) to encode the vectorial token representations into contextualized word representations:

$$\boldsymbol{h}_i = \text{BiLSTM}\left(\boldsymbol{w}_i\right) \quad (2)$$

where $\boldsymbol{h}_i$ is the token hidden representation.

Moreover, in the same way as previous work (Barnes et al., 2021), we also enhance token representations with pretrained contextualized embeddings using multilingual BERT (Devlin et al., 2018).

4.2 Multi-view Token Graph

In this section, we propose a novel multi-view token graph as our hidden layer, which includes four views, span graph, relation graph, [CLS]-related graph and vanilla GAT graph, and each view is full connected with the attention scoring weights as graph edges and the token representations as graph nodes. Recall that the whole label set is applied in this layer, which includes three groups of labels (span, relation and [CLS]-related labels). Thus, three views of graphs (span, relation and [CLS]-related graph) are used to digest information from three groups of labels respectively, while one view (vanilla GAT graph) is not assigned for any specific task, as the method in vanilla graph attention network (GAT) (Veličković et al., 2017). Formally, we represent the latent token graph $\boldsymbol{G}$ as follows:

$$\boldsymbol{G} = \left(\mathbf{V}, S_o^{\mathcal{G}}, S_s^{\mathcal{G}}, S_r^{\mathcal{G}}, S_c^{\mathcal{G}}\right) \tag{3}$$

where superscript $\mathcal{G}$ denotes the graph layer, $\mathbf{V}$ is the set of tokens, $S_o^{\mathcal{G}}$ is the attention scoring matrix in vanilla GAT, $S_s^{\mathcal{G}}$, $S_r^{\mathcal{G}}$ and $S_c^{\mathcal{G}}$ are the attention scoring matrices used to capture information from span, relation and [CLS]-related labels respectively. Without loss of generality, we employ $\mathcal{S}^{\mathcal{G}} = \{S_o^{\mathcal{G}}, S_s^{\mathcal{G}}, S_r^{\mathcal{G}}, S_c^{\mathcal{G}}\}$ unifiedly to represent the four matrices.

4.2.1 Graph Induction

In this section, we introduce the process that we induce the edges of our multi-view token graphs (i.e. four attention scoring matrices $\mathcal{S}^{\mathcal{G}}$) using a mechanism of attention scoring.

Attention Scoring Our attention matrices are produced by a mechanism of attention scoring which takes two token representations $\boldsymbol{h}_i, \boldsymbol{h}_j$ as the input, and for the attention matrix corresponding to a certain view $v \in \{o, s, r, c\}$, we first map the tokens to $q_{v,i}$ and $k_{v,j}$ with two multi-layer perceptions (MLP):

$$q_{v,i}, k_{v,j} = MLP_v^q\left(\boldsymbol{h}_i\right), MLP_v^k\left(\boldsymbol{h}_j\right) \tag{4}$$

Then we apply the technique of Rotary Position Embedding (RoPE) (Su et al., 2021) to encode the relative position information. Thus, for the graph of view v, the attention score $S_{v,ij}^{\mathcal{G}}$ between token i and j can be calculated as follows:

$$S_{v,ij}^{\mathcal{G}} = \left(q_{v,i}\right)^{\top} \boldsymbol{R}_{j-i} k_{v,j} \tag{5}$$

where $\boldsymbol{R}_{j-i}$ can incorporate explicit relative positional information into the attention score $S_{v,ij}^{\mathcal{G}}$. And in the same way as calculating $S_{v,ij}^{\mathcal{G}}$, we can produce the scores of all views and all token pairs, thus inducing the whole graph edges $\mathcal{S}^{\mathcal{G}}$:

$$\mathcal{S}^{\mathcal{G}} = \left\{ S_{v,ij}^{\mathcal{G}} | v \in \{o, s, r, c\}, 1 \leq i, j \leq n \right\} \tag{6}$$

where n is the length of the sentence. The process that the whole label set learnt by attention scoring matrices $S_s^{\mathcal{G}}$, $S_r^{\mathcal{G}}$ and $S_c^{\mathcal{G}}$ through a multi-label adaptive-threshold loss will be introduced in Section 4.4.

4.2.2 Multi-hop Reasoning

Considering that the attention scoring matrix $\mathcal{S}^{\mathcal{G}}$ now fuses rich information, we naturally think of applying a multi-hop reasoning to obtain more informative token representations. Concretely, we first apply a softmax on our adjacency attention matrix $\mathcal{S}^{\mathcal{G}}$, then the computation for the representation $\boldsymbol{u}_i^{l+1}$ of the token i at the $(l+1)^{th}$layer, which takes the representations from previous layer as input and outputs the updated representations, can be defined as:

$$A_v = Softmax\left(S_v^{\mathcal{G}}\right), v \in \{o, s, r, c\} \tag{7}$$

$$\boldsymbol{u}_i^{l+1} = \sigma\left(\frac{1}{N} \sum_{v} \sum_{j \in \mathcal{N}_i^v} A_{v,ij} \boldsymbol{W}_l^v \boldsymbol{u}_j^l\right) \tag{8}$$

where $\boldsymbol{W}_l^v$ is the trainable weight, $\mathcal{N}_i^v$ is the neighbor of token i in graph of view v, σ is the ReLU activation function.

4.3 Adaptive Multi-label Classifier

Considering that the previous sota model (Barnes et al., 2021) is not able to deal with multi-label classification as aforementioned, we propose a novel adaptive multi-label classifier as our prediction layer to identify possible essential labels for each token pair.

Firstly, we take a shortcut connection between the outputs of the encoder layer and graph layer to get the final representation $\boldsymbol{c}_i = \boldsymbol{h}_i \oplus \boldsymbol{u}_i$ for each token. And by taking $\boldsymbol{c}_i$ as the input, we calculate the attention scoring matrices $\mathcal{S}^{\mathcal{P}}$ based on the mechanism of attention scoring (cf. Eq.(4), Eq.(5) and Eq.(6)):

$$\mathcal{S}^{\mathcal{P}} = \{S_r^{\mathcal{P}} | r \in \mathcal{R}_e\} \tag{9}$$

where superscript $\mathcal{P}$ denotes the prediciton layer, $\mathcal{R}_e$ denotes the essential label set. Then, we introduce a technique of adaptive thresholding, which produces a token pair dependent threshold to enable the prediction of the labels for each token pair.

Adaptive Thresholding For a certain token pair with representations of $\boldsymbol{c}_i, \boldsymbol{c}_j$, the token pair dependent threshold $TH_{ij}^{\mathcal{P}}$ and the whole $TH^{\mathcal{P}}$ are calculated as follows:

$$\begin{aligned} TH_{ij}^{\mathcal{P}} &= \left(q_i^{TH}\right)^{\top} \boldsymbol{R}_{j-i} \boldsymbol{k}_j^{TH} \\ TH^{\mathcal{P}} &= \left\{TH_{ij}^{\mathcal{P}} | 1 \leq i, j \leq n\right\} \end{aligned} \tag{10}$$

where $q_i^{TH} = \boldsymbol{W}_q \boldsymbol{h}_i + \boldsymbol{b}_q, k_j^{TH} = \boldsymbol{W}_k \boldsymbol{h}_j + \boldsymbol{b}_k$, the $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{b}_q$ and $\boldsymbol{b}_k$ are the trainable weight and bias matrix, $\boldsymbol{R}_{j-i}$ are calculated in the same way as Eq.(5), which is used to incorporate explicit relative positional information.

Formally, for a certain token pair c_i, c_j, the essential label set is predicted by the following equation:

$$\Omega_{ij} = \left\{r | S_{r,ij}^{\mathcal{P}} > TH_{ij}^{\mathcal{P}}, r \in \mathcal{R}_e\right\} \tag{11}$$

where $\mathcal{R}_e$ denotes the essential label set, Ω_{ij} is the set of predicted labels of token pair c_i, c_j.

4.4 Training

In this section, we will propose a novel loss function, namely multi-label adaptive-threshold loss, to enable a hierarchical training process for our model and our labeling strategy (i.e. whole label set learnt by $S_s^{\mathcal{G}}$, $S_r^{\mathcal{G}}$ and $S_c^{\mathcal{G}}$ in the hidden layer, essential label set learnt by $\mathcal{S}^{\mathcal{P}}$ in the prediction layer), which is based on a variant[2] of Circle loss (Sun et al., 2020), the difference is that we replace the fixed global threshold with the adaptive token pair dependent threshold to enable a flexible and selective learning of more useful information from whole label set.

Take the hidden layer as an example. Actually, we also implement the adaptive thresholding (cf. Eq.(10)) in the hidden layer, where we compute all the token pair dependent threshold $TH^{\mathcal{G}} = \left\{TH_{ij}^{\mathcal{G}} | 1 \leq i, j \leq n\right\}$ by taking the token representation $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ as the input. Then, the multi-label adaptive-threshold loss in hidden layer can be calculated as follows:

$$\begin{aligned} \mathcal{L}_w = &\sum_i \sum_{j>i} \log \left(e^{TH_{ij}^{\mathcal{G}}} + \sum_{r \in \Omega_{ij}^{neg}} e^{S_{r,ij}^{\mathcal{G}}} \right) \\ &+ \sum_i \sum_{j>i} \log \left(e^{-TH_{ij}^{\mathcal{G}}} + \sum_{r \in \Omega_{ij}^{pos}} e^{-S_{r,ij}^{\mathcal{G}}} \right) \end{aligned} \tag{12}$$

where $\Omega_{ij}^{pos} \subseteq \mathcal{R}_w$ and $\Omega_{ij}^{neg} \subseteq \mathcal{R}_w$ are positive and negative classes involving whole labels that exist or not exist between token i and j. When minimizing $\mathcal{L}_w$, the loss pushes the attention score $S_{r,ij}^{\mathcal{G}}$ above the threshold $TH_{ij}^{\mathcal{G}}$ if the token pair possesses the label, while pulls below when it does not.[3]

In a similar way we can calculate the loss $\mathcal{L}_e$ in the prediction layer by taking $TH^{\mathcal{P}}$, $\mathcal{S}^{\mathcal{P}}$ as the inputs of the loss function. Thus the whole loss of our model can be calculated as follows:

$$\mathcal{L}_{all} = \mathcal{L}_e + \alpha \mathcal{L}_w \tag{13}$$

where the α is a hyperparameter to adjust the ratio of the two losses.

5 Experiments

5.1 Datasets and Configuration

For comparison with previous sota work (Barnes et al., 2021), we perform experiments on five structured sentiment datasets in four languages, including multi-domain professional reviews **NoReC**$_{\text{Fine}}$ (Øvrelid et al., 2020) in Norwegian, hotel reviews **MultiB**$_{\text{EU}}$ and **MultiB**$_{\text{CA}}$ (Barnes et al., 2018) in Basque and Catalan respectively, news **MPQA** (Wiebe et al., 2005) in English and reviews of online universities and e-commerce **DS**$_{\text{Unis}}$ (Toprak et al., 2010) in English.

For fair comparison, we use word2vec skip-gram embeddings openly available from the NLPL vector repository [4] (Kutuzov et al., 2017) and enhance token representations with multilingual BERT (Devlin et al., 2018), which has 12 transformer blocks, 12 attention heads, and 768 hidden units. Our network weights are optimized with Adam and we also conduct Cosine Annealing Warm Restarts learning

[2]The variant of Circle loss was proposed by Su on the website https://kexue.fm/archives/7359.

[3]As aforementioned in Section 4.2, three of the attention scoring matrices and three groups of the whole labels have a one-to-one relationship, so here we can index the three matrices with the whole labels.

[4]http://vectors.nlpl.eu/repository.

Dataset	Model	Span				Targeted	Sent. Graph	
		Holder F1	Target F1	Exp. F1	Overall F1	F1	NSF1	SF1
NoReC$_{\text{Fine}}$	RACL-BERT	-	47.2	56.3	-	30.3	-	-
	Head-first	51.1	50.1	54.4	53.1*	30.5	37.0	29.5
	Head-final	60.4	**54.8**	55.5	55.7*	31.9	39.2	31.2
	TGLS	**60.9**	53.2	**61.0**	**58.1**	**38.1**	**46.4**	**37.6**
MultiB$_{\text{EU}}$	RACL-BERT	-	59.9	72.6	-	56.8	-	-
	Head-first	60.4	64.0	73.9	69.6*	57.8	58.0	54.7
	Head-final	60.5	64.0	72.1	68.2*	56.9	58.0	54.7
	TGLS	**62.8**	**65.6**	**75.2**	**71.0**	**60.9**	**61.1**	**58.9**
MultiB$_{\text{CA}}$	RACL-BERT	-	67.5	70.3	-	52.4	-	-
	Head-first	43.0	72.5	71.1	70.5*	55.0	62.0	56.8
	Head-final	37.1	71.2	67.1	70.2*	53.9	59.7	53.7
	TGLS	**47.4**	**73.8**	**71.8**	**71.6**	**60.6**	**64.2**	**59.8**
MPQA	RACL-BERT	-	20.0	31.2	-	17.8	-	-
	Head-first	43.8	51.0	**48.1**	**47.7***	**33.5**	24.5	17.4
	Head-final	**46.3**	49.5	46.0	47.2*	18.6	26.1	18.8
	TGLS	44.1	**51.7**	47.8	47.0	23.3	**28.2**	**21.6**
DS$_{\text{Unis}}$	RACL-BERT	-	44.6	38.2	-	27.3	-	-
	Head-first	28.0	39.9	40.3	40.1*	26.7	31.0	25.0
	Head-final	37.4	42.1	**45.5**	43.0*	29.6	34.3	26.5
	TGLS	**43.7**	**49.0**	42.6	**45.7**	**31.6**	**36.1**	**31.1**

Table 2: Main experimental results of our TGLS model and comparison with previous works. The score marked as bold means the best performance among all the methods. The baseline results with "*" are from our reimplementation, the others are from (Barnes et al., 2021).

rate schedule (Loshchilov and Hutter, 2016). We fixed the word embeddings during training process. The char embedding size is set to 100. The dropout rate of embeddings and other network components are set to 0.4 and 0.3 respectively. We employ 4-layer BiLSTMs with the output size set to 400 and 2-layer for multi-hop reasoning with output size set to 768. The learning rate is 3e-5 and the batch size is 8. The hyperparameter α in Eq.13 is set to 0.25 (cf. Section 6.2). We use GeForce RTX 3090 to train our model for at most 100 epochs and choose the model with the highest SF1 score on the validation set to output results on the test set.

5.2 Baselines

We compare our proposed model with three state-of-the-art baselines which outperform other models in all datasets:

RACL-BERT Chen and Qian (2020) propose a relation-aware collaborative learning framework for end2end sentiment analysis which models the interactive relations between each pair of sub-tasks (target extraction, expression extraction, sentiment classification). Barnes et al. (2021) reimplement the RACL as a baseline for SSA task in their work.

Head-first and Head-final[5] Barnes et al. (2021) cast the structured sentiment analysis as a dependency parsing task and apply a reimplementation of the neural parser by Dozat and Manning (2018), where the main architecture of the model is based on a biaffine classifier. The Head-first and Head final are two models with different setups in the parsing graph.

5.3 Evaluation Metrics

Following previous SOTA work (Barnes et al., 2021), we use the Span F1, Targeted F1 and two Sentiment Graph Metrics to measure the experimental results.

In detail, Span F1 evaluates how well these models are able to identify the holders, targets, and expressions. Targeted F1 requires the exact extraction of the correct target, and the corresponding polarity. Sentiment Graph Metrics include two F1 score, Non-polar Sentiment Graph F1 (NSF1) and Sentiment Graph F1 (SF1), which aims to measure the overall performance of a model to capture the full sentiment graph (Figure 1(a)). For NSF1, each sentiment graph is a tuple of (holder, target, expres-

[5] https://github.com/jerbarnes/sentiment_graphs.

	Span Overall F1	Targeted F1	SF1
Ours(TGLS)	**58.1**	**38.1**	**37.6**
w/o [CLS]-related graph	57.6	36.9	36.1
w/o span graph	57.2	38.1	37.4
w/o relation graph	57.7	38.0	36.1
w/o vanilla GAT graph	57.8	37.6	36.5
w/o RoPE	57.7	36.4	36.8
w/o adaptive thresholding	56.0	36.3	35.2

Table 3: Experimental results of ablation studies.

	$\textbf{NoReC}_{\text{Fine}}$	$\textbf{MultiB}_{\text{EU}}$	$\textbf{MultiB}_{\text{CA}}$	**MPQA**	$\textbf{DS}_{\text{Unis}}$
Head-final	52.3	63.9	67.3	45.0	41.5
TGLS model					
+parsing labels	54.2	65.4	67.5	44.7	43.2
+our labels	**57.8**	**68.7**	**70.1**	**46.1**	**45.7**

Table 4: Experimental results of the relation extraction F1 score, where *parsing labels* denote the dependency-parsing-based labels in head-final setting, *our labels* denote the whole and essential labels.

sion), while SF1 adds the polarity (holder, target, expression, polarity). A true positive is defined as an exact match at graph-level, weighting the overlap in predicted and gold spans for each element, averaged across all three spans.

Moreover, for ease of analysis, we add an Overall Span F1 score which evaluates how well these models are able to identify all three elements of a sentiment graph with token-level F1 score.

5.4 Main Results

In this section, we introduce the main experimental results compared with three state-of-the-art models RACL-BERT (Chen and Qian, 2020), Head-first and Head-final models (Barnes et al., 2021).

Table 2 shows that in most cases our model performs better than other baselines in terms of the Span F1 metrics across all datasets. The average improvement (↑ 1.4) in Overall Span F1 score proves the effectiveness of our model in span extraction. Besides, there exists some significant improvements such as extracting holder on $\textbf{DS}_{\text{Unis}}$ (↑6.3) and extracting expression on $\textbf{NoReC}_{\text{Fine}}$ (↑4.7), but the extracting expression on $\textbf{DS}_{\text{Unis}}$ (↓2.9) are poor.

As for the metric of Targeted F1, although the Head-first model performs well on **MPQA**, our TGLS model is obviously more robust as we achieves superior performance on other 4 datasets. There are also extremely significant improvements such as on $\textbf{NoReC}_{\text{Fine}}$ (↑6.2) and on $\textbf{MultiB}_{\text{CA}}$ (↑5.6), it proves the capacity of our model in exact prediction of target and the corresponding polar.

As for the Sentiment Graph metrics, which

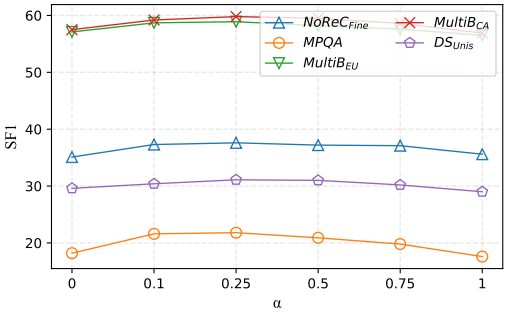

Figure 4: Experimental results (SF1 score) using different α to control the impact of the whole label prediction.

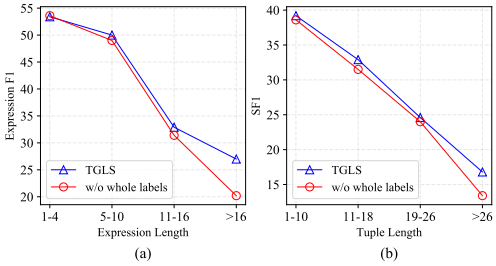

Figure 5: Analysis on the effect of the whole label set for long span identification. **(a)** Expression F1 scores regarding to different expression lengths. **(b)** SF1 scores regarding to different tuple lengths.

are important for comprehensively examining span, relation and polar predictions, our TGLS model achieves superior performance throughout all datasets in both NSF1 and SF1 score, especially on $\textbf{NoReC}_{\text{Fine}}$ (↑7.2 and ↑6.4). And the average improvement (↑4.5) in SF1 score verifies the excellent ability of our model in the end-to-end sentiment tuple extraction.

5.5 Ablation Study

In this section, we conduct extensive ablation studies on $\textbf{NoReC}_{\text{Fine}}$ to better understand independent contributions of different components in terms of span overall F1, targeted F1 and SF1 scores.

Firstly, we remove each view of our graphs separately. As shown in Table 3, we observe that the [CLS]-related graph is effective in all three metrics which proves the importance of utilizing sentence-level global information. As we assumed, the span graph makes more contribution to the performance of span extraction (Span Overall F1) while the relation graph contributes more to end-to-end sentiment tuple extraction (SF1). And we also observe that the vanilla GAT graph makes consid-

erable improvement in SF1 score.

Then, we test the effectiveness of the Rotary Position Embedding (RoPE) (Su et al., 2021). The results in Table 3 demonstrate that RoPE can make our model more sensitive to the relative positional information since it significantly improves the performance of exact target extraction (Targeted F1).

Last, we replace the adaptive threshold with fixed global threshold, and we observe that the performance drops drastically in all three metrics, it suggests that the adaptive thresholding mechanism is very crucial for our model since the flexibility can allow our model to selectively learn more useful information for SSA task from whole labels.

6 Analysis

In this section we perform a deeper analysis on the models in order to answer three research questions:

6.1 Does our modeling strategy mitigate the label imbalance problem in span prediction and span relation prediction?

Experimental results in Table 2 show that our model performs significantly better in the SF1 score, which to some extent proves that our model can ensure the efficiency of relation extraction. However, there lacks a metric to directly quantify the ability in relation extraction and it is still a worthy question to explore how much of the improvement comes from our new model and how much from our new labeling strategy?

To answer the question, we replace our labels with the dependency-parsing-based labels in head-final setting (Barnes et al., 2021) and experiment on all datasets in terms of a new relation prediction metric, where a true positive is defined as any span pair that overlaps the gold span pair and has the same relation. Table 4 shows that our new model achieves superior performance of relation prediction than the previous sota model (Barnes et al., 2021). Besides, with new labeling strategy, we can see that our model significantly improve the performance on all datasets compared with the model with replaced dependency-parsing-based labels.

6.2 What is the appropriate value for the hyperparameter α in Eq. 13?

In this section, we experiment on five datasets to heuristically search for the appropriate value of hyperparameter α (cf. Eq.(13)). Figure 4 shows that all datasets achieve higher SF1 score with α between 0.1 and 0.5. We ended up fixing alpha to 0.25, since most datasets yield optimal results around this value. In addition, it is worth noting that when α is set to 0, which means that the whole labels are completely removed, the performance drops a lot, which once again proves the effectiveness of learning whole labels in the hidden layer.

6.3 Is the whole label set helpful for long span identification?

In this section, we experiment on $\textbf{NoReC}_{\text{Fine}}$ to further explore whether whole labels contribute to long span identification. Figure 5(a) evaluates the Expression F1 scores regarding to different expression lengths, we can find that whole labels helps most on those expressions with longer length. In Figure 5(b), we also report the SF1 scores regarding to different distances, that is, from the leftmost token in a tuple to the rightmost token, which shows a similar conclusion.

7 Conclusion

In this paper, we propose a token graph model with a novel labeling strategy, consisting of the whole and essential label sets, to extract sentiment tuples for structured sentiment analysis. Our model is capable of modeling both global and local token pair interactions by jointly predicting whole labels in the hidden layer and essential labels in the output layer. More importantly, our modeling strategy is able to alleviate the label imbalance problem when using token-graph-based approaches for SSA. Experimental results show that our model overwhelmingly outperforms SOTA baselines and improves the performance of identifying the sentiment components with long spans. We believe that our labeling strategy and model can be well extended to other structured prediction tasks.

Acknowledgements

We thank all the reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (No. 62176187), the National Key Research and Development Program of China (No. 2017YFC1200500), the Research Foundation of Ministry of Education of China (No. 18JZD015), the Youth Fund for Humanities and Social Science Research of Ministry of Education of China (No. 22YJCZH064), the General Project of Natural Science Foundation of Hubei Province (No.2021CFB385).

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 LREC*, 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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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 ACL*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

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.

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 271–276. Linköping University Electronic Press.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th LREC*, pages 5025–5033, Marseille, France. European Language Resources Association.

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.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6397–6406.

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.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

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 EMNLP*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

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 NAACL*, pages 1795–1804.

Meishan Zhang, Qiansheng Wang, and Guohong Fu. 2019. End-to-end neural opinion extraction with a transition-based model. *Information Systems*, 80:56–63.