

Improving Aspect Sentiment Quad Prediction via Template-Order Data Augmentation

Mengting Hu¹ Yike Wu² Hang Gao^{3*} Yin hao Bai¹ Shiwan Zhao[†]

¹ College of Software, Nankai University

² School of Journalism and Communication, Nankai University

³ Institute for Public Safety Research, Tsinghua University

{mthu, wuyike}@nankai.edu.cn, gaohang@mail.tsinghua.edu.cn

yinhao@mail.nankai.edu.cn, zhaosw@gmail.com

Abstract

Recently, aspect sentiment quad prediction (ASQP) has become a popular task in the field of aspect-level sentiment analysis. Previous work utilizes a predefined template to paraphrase the original sentence into a structure target sequence, which can be easily decoded as quadruplets of the form (*aspect category, aspect term, opinion term, sentiment polarity*). The template involves the four elements in a fixed order. However, we observe that this solution contradicts with the order-free property of the ASQP task, since there is no need to fix the template order as long as the quadruplet is extracted correctly. Inspired by the observation, we study the effects of template orders and find that some orders help the generative model achieve better performance. It is hypothesized that different orders provide various views of the quadruplet. Therefore, we propose a simple but effective method to identify the most proper orders, and further combine multiple proper templates as data augmentation to improve the ASQP task. Specifically, we use the pre-trained language model to select the orders with minimal entropy. By fine-tuning the pre-trained language model with these template orders, our approach improves the performance of quad prediction, and outperforms state-of-the-art methods significantly in low-resource settings¹.

1 Introduction

The aspect sentiment quad prediction (ASQP) task, aiming to extract aspect quadruplets from a review sentence, becomes popular recently (Zhang et al., 2021a; Cai et al., 2021). The quadruplet consists of four sentiment elements: 1) *aspect category* (*ac*) indicating the aspect class; 2) *aspect term* (*at*) which is the specific aspect description; 3) *opinion term*

Original sentence	<i>The restaurant is clean.</i>
Quadruplet (<i>ac, at, ot, sp</i>)	(<i>ambience general, restaurant, clean, positive</i>)
Semantic quadruplet ($x_{ac}, x_{at}, x_{ot}, x_{sp}$)	(<i>ambience general, restaurant, clean, great</i>)
Fixed-order template	x_{ac} is x_{sp} because x_{at} is x_{ot}
Target sequence	<i>ambience general is great because restaurant is clean</i>
Free-order template	$O_i([AC] x_{ac} [AT] x_{at} [OT] x_{ot} [SP] x_{sp}); i \in [1, 24]$
Multiple target sequences	[AC] x_{ac} [AT] x_{at} [OT] x_{ot} [SP] x_{sp} [AT] x_{at} [AC] x_{ac} [OT] x_{ot} [SP] x_{sp} ...

Figure 1: An example sentence is paraphrased into a target sequence with a fixed-order template (Zhang et al., 2021a). Our approach employs special markers to form free-order templates and generates multiple target sequences. O_i is the i -th order permutation of the four elements.

(*ot*) which is the opinion expression towards the aspect; 4) *sentiment polarity* (*sp*) denoting the sentiment class of the aspect. For example, the sentence “*The service is good and the restaurant is clean.*” contains two quadruplets (*service general, service, good, positive*) and (*ambience general, restaurant, clean, positive*).

To extract aspect sentiment quadruplets, Zhang et al. (2021a) propose a new paradigm which transforms the quadruplet extraction into paraphrase generation problem. With pre-defined rules, they first map the four elements of (*ac, at, ot, sp*) into semantic values ($x_{ac}, x_{at}, x_{ot}, x_{sp}$), which are then fed into a template to obtain a nature language target sequence. As shown in Figure 1, the original sentence is “re-written” into a target sequence by paraphrasing. After fine-tuning the pre-trained language model (Raffel et al., 2020) in such a sequence-to-sequence learning manner, the quadruplets can be disentangled from the target sequence.

Though promising is this paradigm, one issue is that the decoder of the generative pre-trained language model (Raffel et al., 2020) is unidirectional (Vinyals et al., 2015), which outputs the target se-

*Corresponding author.

†Independent researcher.

¹Experimental codes and data are available at: <https://github.com/hmt2014/AspectQuad>.

quence from the beginning of the sequence to its end. Thus four elements of a quadruplet are modeled in a fixed order $\{x_{ac} \rightarrow x_{sp} \rightarrow x_{at} \rightarrow x_{ot}\}$. Yet ASQP is not a typical generation task. There is no need to fix the element order of the quadruplet as long as it can be extracted accurately. Aspect sentiment quadruplet has the order-free property, suggesting that various orders, such as $\{x_{ac} \rightarrow x_{sp} \rightarrow x_{at} \rightarrow x_{ot}\}$ and $\{x_{at} \rightarrow x_{ac} \rightarrow x_{ot} \rightarrow x_{sp}\}$, are all correct.

In light of this observation, our curiosity is triggered: Does the order of the four elements impact the generative pre-trained language models’ performances? Thus we conduct a pilot experiment. The four elements are concatenated with commas, thus we could switch their orders in a flexible manner and obtain order permutations. It is found that some template orders can help the generative model perform better. Even only concatenating with commas, some orders outperform the state-of-the-art.

It is hypothesized that different orders provide various views of the quadruplet. Therefore, we propose a simple but effective method to identify the most proper orders, and further combine multiple proper templates as data augmentation to improve the ASQP task. Concretely, we use the pre-trained language model (Raffel et al., 2020) to select the orders with minimal entropy. Such template orders can better promote the potential of the pre-trained language model. To jointly fine-tune these template orders together, inspired by Paolini et al. (2021), we design special markers for the four elements, respectively. The markers help to disentangle quadruplets by recognizing both the types and their values of the four elements from the target sequence. In this way, the template orders do not need to be fixed in advance.

In summary, the contributions of this work are three-fold:

- We study the effects of template orders in the ASQP task, showing that some orders perform better. To the best of our knowledge, this work is the first attempt to investigate ASQP from the template order perspective.
- We propose to select proper template orders by minimal entropy computed with pre-trained language models. The selected orders are roughly consistent with their ground-truth performances.
- Based on the order-free property of the

quadruplet, we further combine multiple proper templates as data augmentation to improve the ASQP task. Experimental results demonstrate that our approach outperforms state-of-the-art methods and has significant gains in low-resource settings.

2 Preliminaries on Generative ASQP

2.1 Paraphrase Generation

Given a sentence x , aspect sentiment quad prediction (ASQP) aims to extract all aspect-level quadruplets $\{(ac, at, ot, sp)\}$. Recent paradigm for ASQP (Zhang et al., 2021a) formulates this task as a paraphrase generation problem. They first define projection functions to map quadruplet (ac, sp, at, ot) into semantic values $(x_{ac}, x_{sp}, x_{at}, x_{ot})$. Concretely, 1) aspect category ac is transformed into words, such as $x_{ac} = \text{“service general”}$ for $ac = \text{“service#general”}$; 2) if aspect term at is explicit, $x_{at} = at$, otherwise $x_{at} = \text{“it”}$; 3) if opinion term ot are explicitly mentioned, $x_{ot} = ot$, otherwise it is mapped as “NULL” if being implicitly expressed; 4) the sentiment polarity $sp \in \{\text{positive, neutral, negative}\}$, is mapped into words with sentiment semantics $\{\text{great, ok, bad}\}$, respectively.

With the above rules, the values can better exploit the semantic knowledge from pre-trained language model. Then the values of quadruplet are fed into the template, which follows the cause and effect semantic relationship.

$$x_{ac} \text{ is } x_{sp} \text{ because } x_{at} \text{ is } x_{ot}. \quad (1)$$

It is worth noting that if a sentence describes multiple quadruplets, the paraphrases are concatenated with a special marker [SSEP] to obtain the final target sequence y .

2.2 Sequence-to-Sequence Learning

The purpose of paraphrasing is consistent with the typical sequence-to-sequence problem. The encoder-decoder model is leveraged to *“re-write”* the original sentence x into the target sequence y . Assume the parameter is θ , the overall objective is to model the conditional probability $p_{\theta}(y|x)$. Specifically, at the t -th time step, the decoder output y_t is calculated with the input x and the previous outputs $y_{<t}$, formulating as below.

$$p_{\theta}(y_{t+1}|x, y_{<t+1}) = \text{softmax}(W^T y_t) \quad (2)$$

where W maps y_t into a vector, which can represent the probability distribution over the whole vocabulary set.

Target Sequence	Rest15		
	Pre	Rec	F1
$x_{sp}, x_{ac}, x_{at}, x_{ot}$	45.55	46.34	45.94
$x_{sp}, x_{ot}, x_{at}, x_{ac}$	46.12	47.52	46.81
$x_{ac}, x_{ot}, x_{at}, x_{sp}$	47.07	47.85	47.46
$x_{ac}, x_{sp}, x_{ot}, x_{at}$	47.60	48.75	48.17

Target Sequence	Rest16		
	Pre	Rec	F1
$x_{ot}, x_{at}, x_{sp}, x_{ac}$	56.04	58.17	57.09
$x_{ac}, x_{ot}, x_{sp}, x_{at}$	57.14	58.72	57.92
$x_{at}, x_{sp}, x_{ac}, x_{ot}$	57.35	59.60	58.45
$x_{ac}, x_{sp}, x_{at}, x_{ot}$	58.11	60.33	59.20

Table 1: Evaluation results on various template orders in terms of precision (Pre, %), recall (Rec, %) and F1 (F1, %) scores. All the reported results are the average of five runs. Full results are shown in the appendix, where some template orders outperform Paraphrase.

During training, a pre-trained encoder-decoder model, i.e. T5 (Raffel et al., 2020), is chosen to initialize the parameter θ and fine-tuned with minimizing the cross-entropy loss.

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = - \sum_{t=1}^n \log p_{\theta}(\mathbf{y}_t | \mathbf{x}, \mathbf{y}_{<t}) \quad (3)$$

where n is the length of the target sequence \mathbf{y} .

3 A Pilot Experiment

As Eq. (1) displayed, this template forms a fixed order of four elements. Our curiosity is whether the quadruplet’s order affects the performance of sequence-to-sequence learning. Therefore, we conduct a pre-experiment. By only concatenating with commas, four elements can also be transformed into a target sequence. The orders can be switched in a more flexible way, compared with Eq. (1). There will be $4! = 24$ permutations. During inference, quadruplets can be recovered by splitting them with commas. Based on the pre-experimental results, we have the following observations.

Template order affects the performances of sequence-to-sequence learning. Part of the experimental results on two datasets, i.e. Rest15 and Rest16, are shown in Table 1. It is observed that on Rest15, the F1 score ranges from 45.94% to 48.17%. Similarly, on Rest16, the F1 score ranges from 57.09% to 59.20%. We draw an empirical conclusion that the template order also matters for the ASQP task. Moreover, a template has various performances on different datasets, which is hard to say some order is absolutely good.

		Position						Position			
		1	2	3	4			1	2	3	4
Elements	x_{ac}	86.16	85.10	84.01	83.28	Elements	x_{ac}	88.58	88.30	88.48	88.32
	x_{at}	80.15	80.40	80.39	80.52		x_{at}	86.12	85.70	86.12	86.25
	x_{ot}	68.41	69.83	70.06	69.05		x_{ot}	77.27	78.21	78.63	78.52
	x_{sp}	92.45	91.75	90.64	89.44		x_{sp}	94.20	93.98	93.52	93.06

(a) Results on Rest15

(b) Results on Rest16

Figure 2: Evaluation results of each element at each position in terms of F1 score (%). The performances of different elements show different trends with the position. The best score of each row is marked in bold.

The performances of each element are connected to its position. We further investigate the F1 scores on each of the four elements. Given the 24 permutations, there are 6 templates for each element at each position. For example, there are 6 templates $\{(\cdot, \cdot, x_{ac}, \cdot)\}$ of x_{ac} at position 3. In Figure 2, we show the average F1 scores of the six templates of each element at each position. We can see that the performances of the four elements have different trends with the positions. The F1 scores of x_{ac} and x_{sp} both degrade when they are gradually placed backwards. Compared with the other three elements, x_{at} is more stable on different positions while x_{ot} has the worst performance in the first position. In addition to positions, it can also be observed that x_{ac} and x_{sp} achieve higher F1 scores compared with x_{at} and x_{ot} , showing various difficult extents of four elements.

4 Methodology

As analyzed in the previous section, the template order influences the performances of both the quadruplet and its four elements. It is hypothesized that different orders provide various views of the quadruplet. We argue that combining multiple template orders may improve the ASQP task via data augmentation. However, using all 24 permutations significantly increases the training time, which is inefficient. Therefore, we propose a simple method to select proper template orders by leveraging the nature of the pre-trained language model (i.e. T5). Then for the ASQP task, these selected orders are utilized to construct the target sequence \mathbf{y} to fine-tune the T5 model.

Specifically, given an input sentence \mathbf{x} and its quadruplets $\{(ac, at, ot, sp)\}$, following Zhang

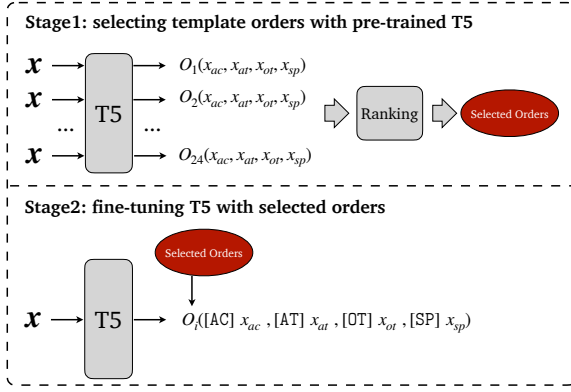


Figure 3: The proposed method is composed of two stages. The first stage aims to select template orders via pre-trained T5. The second stage constructs training samples with the selected orders and fine-tunes T5.

et al. (2021a), we map them into semantic values $\{(x_{ac}, x_{at}, x_{ot}, x_{sp})\}$. As shown in Figure 3, our approach is composed of two stages, next which will be introduced in detail.

4.1 Selecting Template Orders

Inspired by (Yuan et al., 2021; Lu et al., 2022), we choose template orders by evaluating them with the pre-trained T5. As shown in Figure 3, given an input x and its quadruplets, we construct all 24 target sequences with multiple order mapping functions O_i , where $i \in [1, 24]$. An example O_i is shown below.

$$O_i(x_{ac}, x_{at}, x_{ot}, x_{sp}) = x_{at} x_{ac} x_{ot} x_{sp} \quad (4)$$

where the four values are concatenated with a simple space, without any other tokens such as commas, in a specific order O_i . In this way we can reduce the impact of noisy tokens, but focus more on the order. We also introduce the special symbol [SSEP] if there are multiple quadruplets in a sentence. Given multiple template orders, multiple target sequences y_{o_i} are constructed for an input x .

Then we evaluate these target sequences with the entropy computed by the pre-trained T5. Here y_{o_i} is also fed into the decoder as teacher forcing (Williams and Zipser, 1989). The output logits p_θ of the decoder are utilized to compute the entropy.

$$\mathcal{E}(y_{o_i} | x) = -\frac{1}{n} \sum_n \sum_{|V|} p_\theta \log p_\theta \quad (5)$$

where n is length of target sequence and $|V|$ is the size of the vocabulary set.

Given the whole training set \mathcal{T} , we have $(x, \{y_{o_i}\}_{i=1}^{24})$ for each instance by constructing

template orders. Specifically, we design the following two template selection strategies.

Dataset-Level Order (DLO) To choose the dataset-level orders, we compute a score for each order on the whole training set.

$$\mathcal{S}_{o_i} = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}} \mathcal{E}(y_{o_i} | x) \quad (6)$$

where \mathcal{S}_{o_i} denotes the average entropy of all instances for the template order O_i . Then by ranking these scores, template orders with smaller values are chosen.

Instance-Level Order (ILO) Different instances have various contexts and semantics, and tend to have their own proper template orders. Therefore, we also design to choose orders at the instance level. Similarly, the template orders of each instance with small values are chosen based on Eq. (5).

4.2 Fine-tuning with Selected Orders

Multiple template orders provide various views of a quadruplet. However, to train them jointly, an issue arises. If the four values are concatenated with a comma or only a blank space, the value type could not be identified during the inference. For example, when the output sequence “*food quality, pasta, delicious, great*” is recovered to a quadruplet, the machine does not know the element types. Therefore, to deal with this issue, we design special markers to represent the structure of the information (Paolini et al., 2021). The markers for x_{ac} , x_{at} , x_{ot} , x_{sp} are [AC], [AT], [OT], [SP], respectively. Given an order, the target sequence is constructed:

$$\begin{aligned} y_{o_i} &= O_i([\text{AC}] x_{ac}, [\text{AT}] x_{at}, [\text{OT}] x_{ot}, [\text{SP}] x_{sp}) \\ &= [\text{AT}] x_{at} [\text{AC}] x_{ac} [\text{SP}] x_{sp} [\text{OT}] x_{ot} \end{aligned}$$

Now we can train multiple orders together, meanwhile the quadruplet can be recovered by these special markers during the inferences. Note that previous data augmentation methods usually design multiple inputs for one label, such as word deletion and replacement (Gao et al., 2021), obtaining multiple input sentences. While our method constructs multiple labels for one input sequence. This is beneficial from the ASQP task’s property when using generation-based models.

5 Experiments

5.1 Datasets

We conduct experiments on two public datasets, i.e. Rest15 and Rest16 (Zhang et al., 2021a).

Methods	Rest15			Rest16		
	Pre	Rec	F1	Pre	Rec	F1
HGCN-BERT+BERT-Linear* (Cai et al., 2020)	24.43	20.25	22.15	25.36	24.03	24.68
HGCN-BERT+BERT-TFM* (Cai et al., 2020)	25.55	22.01	23.65	27.40	26.41	26.90
TASO-BERT-Linear* (Wan et al., 2020)	41.86	26.50	32.46	49.73	40.70	44.77
TASO-BERT-CRF* (Wan et al., 2020)	44.24	28.66	34.78	48.65	39.68	43.71
Extract-Classify-ACOS (Cai et al., 2021)	35.64	37.25	36.42	38.40	50.93	43.77
GAS* (Zhang et al., 2021b)	45.31	46.70	45.98	54.54	57.62	56.04
Paraphrase* (Zhang et al., 2021a)	46.16	47.72	46.93	56.63	59.30	57.93
DLO	47.08	49.33	48.18	57.92	61.80	59.79
ILO	47.78	50.38	49.05	57.58	61.17	59.32

Table 2: Evaluation results compared with baseline methods in terms of precision (Pre, %), recall (Rec, %) and F1 score (F1, %). The best scores are marked in bold. The experimental results of baseline methods, marked with *, are obtained from this work (Zhang et al., 2021a).

	Rest15				Rest16			
	#S	#+	#0	#-	#S	#+	#0	#-
Train	834	1005	34	315	1264	1369	62	558
Dev	209	252	14	81	316	341	23	143
Test	537	453	37	305	544	583	40	176

Table 3: Data statistics. #S, #+, #0, and #- denote the number of sentences, the number of positive, neutral and negative quads respectively.

These two datasets originate from SemEval tasks (Pontiki et al., 2015, 2016), which are gradually annotated by previous researchers (Peng et al., 2020; Wan et al., 2020). After alignment and completion by Zhang et al. (2021a), each instance in the two datasets contains a review sentence, with one or multiple sentiment quadruplets. The statistics are presented in Table 3.

5.2 Implementation Details

We adopt T5-base (Raffel et al., 2020) as the pre-trained generative model. The pre-trained parameters are utilized to initialize the model, which is exploited to calculate template orders' entropy without updating any parameters. After selecting order with minimal entropy, we fine-tune T5 with the constructed training samples. The batch size is set to 16. In the pilot experiment, the hyper-parameters are set following Zhang et al. (2021a). The learning rate is set to $3e-4$. During the inference, greedy decoding is chosen to generate the output sequence. The number of training epochs is 20 for all experiments. For the proposed approaches, since multiple template orders are combined, we set the learning rate as $1e-4$ to prevent overfitting. During the inference, we utilize the beam search decoding, with the number of beam being 5, for generating the output sequence. All reported results are the average of 5

fixed seeds.

5.3 Compared Methods

To make an extensive evaluation, we choose the following strong baseline methods.

- **HGCN-BERT+BERT-Linear** HGCN (Cai et al., 2020) aims to jointly extract *ac* and *sp*. Following it, BERT extracts *at* and *ot* (Li et al., 2019). Finally, stacking a linear layer (BERT-Linear) forms the full model.
- **HGCN-BERT+BERT-TFM** The final stacked layer in the above model is changed to a transformer block (BERT-TFM).
- **TASO-BERT-Linear** TAS (Wan et al., 2020) is proposed to extract (*ac*, *at*, *sp*) triplets. By changing the tagging schema, it is expanded into TASO (TAS with Opinion). Followed by a linear classification layer, the model is named as TASO-BERT-Linear.
- **TASO-BERT-CRF** TASO is followed with a CRF layer, named as TASO-BERT-CRF.
- **Extract-Classify-ACOS** (Cai et al., 2021) It is a two-stage method, which first extracts *at* and *ot* from the original sentence. Based on it, *ac* and *sp* are obtained through classification.
- **GAS** (Zhang et al., 2021b) It is the first work to deal aspect level sentiment analysis with generative method, which is modified to directly treat the sentiment quads sequence as the target sequence.
- **Paraphrase** (Zhang et al., 2021a) It is also a generation-based method. By paraphrasing the original sentence, the semantic knowledge from

Methods	Rest15		
	Pre	Rec	F1
ILO(Entropy Max)	45.68	48.85	47.21
ILO(random)	46.84	50.33	48.52
ILO	47.78	50.38	49.05

Methods	Rest16		
	Pre	Rec	F1
DLO(Entropy Max)	57.09	60.77	58.87
DLO(random)	56.92	60.91	58.85
DLO	57.92	61.80	59.79

Table 4: Evaluation results of ablation study.

the pre-trained language model can be better exploited.

5.4 Experimental Results

5.4.1 Overall Results

Experimental results of various approaches are reported in Table 2. The best scores on each metric are marked in bold. It is worth noting that for our two approaches, i.e. ILO and DLO, the default template orders are selected with the top-3 minimal entropy from all permutations.

We observe that ILO and DLO achieve the best performances compared with strong baselines. Specifically, comparing with Paraphrase, the absolute improvement of ILO is +2.12% (+4.51% relatively) F1 score on Rest15 dataset. ILO outperforms Paraphrase by +1.86% (+3.21% relatively) F1 score on Rest16 dataset. This validates the effectiveness of our template-order data augmentation, which provides more informative views for pre-trained models. Our method exploits the order-free property of quadruplet to augment the “output” of a model, which is different from the previous data augmentation approaches.

5.4.2 Ablation Study

To further investigate the strategies for selecting template orders, an ablation study is conducted. The results are shown in Table 4. As aforementioned, the default setting of ILO and DLO is to select top-3 template orders with minimal entropy. The model variants also select top-3 template orders, but with maximal entropy and random sampling. We observe that using minimal entropy consistently outperforms the other two strategies. This verifies that our strategy is effective, and the selected template orders can better promote the potential of T5 on solving the ASQP task.

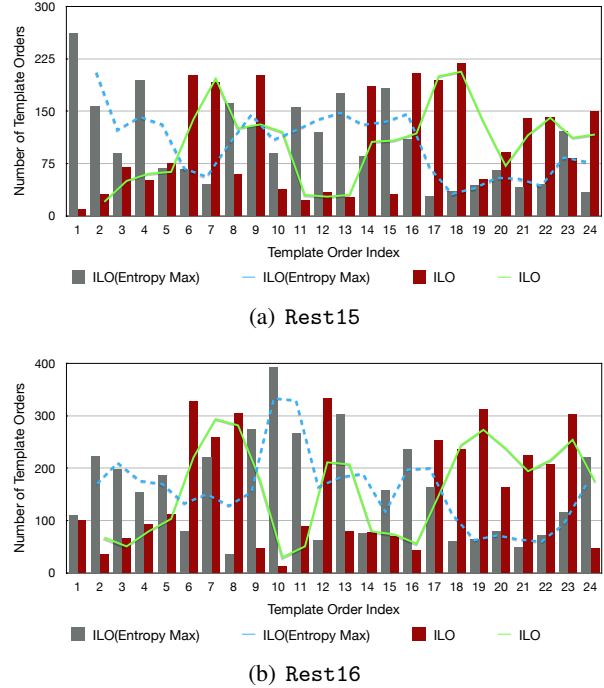


Figure 4: The distribution of the selected templates by ILO and ILO(Entropy Max), respectively, on two datasets. The two curves are the moving average.

Moreover, we investigate the distribution of the chosen template orders. Firstly, we sort all the 24 template orders by their F1 scores in ascending order based on the results of the pilot experiment (see Appendix). As depicted in Figure 4, the horizontal axis represents the template index $i \in [1, 24]$. The template order at index 1 has the worst performance while index 24 the best. We then count the number of each template index which is selected by ILO. We observe that by minimal entropy, more performant template orders (e.g., index $i \in [17, 24]$) are selected compared with using maximal entropy. On the contrary, we also see that ILO chooses less poorly-performed template orders (e.g., index $i \in [1, 5]$) than ILO(Entropy Max). This verifies that using minimal entropy, we can select performant template orders. The observations are similar in DLO, which are presented in the appendix.

5.4.3 Low-Resource Scenario

To further explore the performances of the proposed method in low-resource settings, we design an experiment which uses only 25% of the original training data to train the model. The experimental results are shown in Table 5. Our approach achieves significant improvements compared with the state-of-the-art. Specifically, ILO(top-10) out-

Methods	Rest15		
	Pre	Rec	F1
Paraphrase	35.52	37.76	36.60
ILO	35.66	41.05	38.16
ILO(top-10)	39.12	43.27	41.08

Methods	Rest16		
	Pre	Rec	F1
Paraphrase	47.87	48.96	48.40
ILO	48.51	53.66	50.96
ILO(top-10)	51.31	54.29	52.76

Table 5: Evaluation results in the low-resource scenario, where we only exploit 25% of the training data on two datasets.

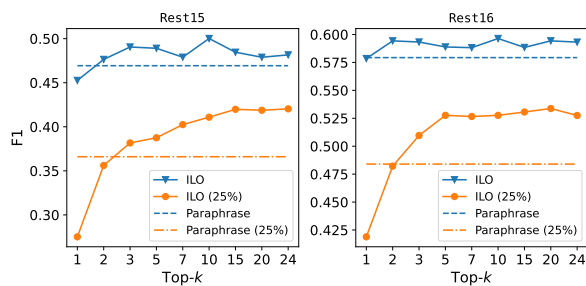


Figure 5: F1 scores of ILO variants by setting different top- k . It is worth noting that (25%) indicates the low-resource scenario, which uses only 25% training data.

performs Paraphrase by +4.48% (+12.24% relatively) and +4.36% (+9.01% relatively) F1 scores on Rest15 and Rest16, respectively. This further verifies our hypothesis that different orders provide various informative views of the quadruplet while combining multiple orders as data augmentation can improve the model training, especially in low-resource settings.

We also plot the F1 score curves by setting different top- k values (see Figure 5). It can be seen that under the two settings, i.e. full and 25% training data, ILO both outperforms Paraphrase. Comparing the two settings, ILO achieves more significant improvements under the low-resource scenario. This observation is in line of expectation. When the training data is adequate, selecting template orders with top-3 is enough. When the training data is limited, model can obtain more gains by setting large k . It also shows that our data augmentation is friendly for real applications which have limited labeled data.

5.4.4 Effects of Special Marker

Since we design four special markers for the four elements to jointly train models with multiple tem-

Template	Rest15			Rest16		
	Pre	Rec	F1	Pre	Rec	F1
T1	48.46	49.46	48.95	58.27	60.33	59.28
T2	47.11	48.30	47.69	57.77	60.40	59.05
T3	47.62	48.60	48.10	57.09	59.57	58.20
T4	46.67	47.87	47.26	57.66	59.97	58.79

Table 6: Evaluation results of special marker.

plates, we investigate the differences in using other symbols. The templates below are chosen for comparison. **T2** and **T3** are inspired by Chia et al. (2022), which annotate the type of information by specific words.

- **T1**: [AT] x_{at} [OT] x_{ot} [AC] x_{ac} [SP] x_{sp}
- **T2**: aspect term: x_{at} opinion term: x_{ot} aspect category: x_{ac} sentiment polarity: x_{sp}
- **T3**: Aspect Term: x_{at} Opinion Term: x_{ot} Aspect Category: x_{ac} Sentiment Polarity: x_{sp}
- **T4**: $x_{at}, x_{ot}, x_{ac}, x_{sp}$

The evaluation results of the above four templates are reported in Table 6. Firstly, by comparing **T4** with others, it can be seen that marking the types of four elements are effective for generative ASQP. A possible reason is that marking with either special symbols or specific words helps to demonstrate the structured information (Paolini et al., 2021). Secondly, **T1** achieves the best performances on almost all evaluation metrics. Such special markers can avoid overlapping words with sentences. For example, the sentence “Service is not what one would expect from a joint in this price category.” contains the word “category”, which is overlapped with the type indicator *aspect category* in **T2**. It is shared between sentence and type markers through word embeddings, and might lead to negative effects.

5.4.5 Error Analysis

We further investigate some error cases. Two example cases are presented in Figure 6. We observe that ILO can generate quadruplets in different orders with the help of special markers. By recognizing the special markers, the quadruplets can be disentangled from the target sequence.

The two examples demonstrate that some cases are still difficult for our approach. The first example contains an implicit aspect term, which is

Example 1
Sentence: <i>Save your money and your time and go somewhere else.</i>
Gold label: [AT] it [OT] go somewhere else [AC] restaurant general [SP] bad
Prediction: [OT] Save [AT] it [AC] great [SP] restaurant miscellaneous ✘
Example 2
Sentence: <i>Terrible, Waste of money .. scammers</i>
Gold label: [AT] it [OT] Terrible [AC] restaurant general [SP] bad [SSEP] [AT] it [OT] Waste of money [AC] restaurant prices [SP] bad
Prediction: [OT] Terrible [SP] bad [AC] service general [AT] it ✘

Figure 6: Two error examples predicted by ILO. It is worth noting that the gold label only provides the value of four elements, but does not constrain the order.

mapped into “it”. Its opinion term, i.e. “go somewhere else”, also expresses negative sentiment polarity implicitly. This case is wrongly predicted. As for the second one, its gold label consists of two quadruplets. Our method only predicts one quadruplet, which does not match either quadruplet. This example also describes aspect terms implicitly for different aspect categories, i.e. “restaurant general” and “restaurant miscellaneous”. In summary, sentences with implicit expressions and multiple aspects are usually tough cases. This observation is also consistent with the results from the pilot experiment. As shown in Figure 2, the F1 scores of aspect term and opinion term are much worse than other two elements.

6 Related Work

6.1 Aspect-Level Sentiment Analysis

Aspect-level sentiment analysis presents a research trend that deals with four elements gradually in a finer-grained manner (Zhang et al., 2022). Analyzing sentiment at the aspect level begins from learning the elements separately (Pontiki et al., 2014). To name a few, some works have been proposed to classify sentiment polarity given the mentioned aspect, either aspect category (Hu et al., 2019) or aspect term (Zhang and Qian, 2020). Other works extract aspect term (Ma et al., 2019), classify aspect category (Bu et al., 2021). The four elements are not solely existing, which actually have strong connections with each other. Therefore, researchers focus on learning them jointly, such as aspect sentiment pair (Zhao et al., 2020; Cai et al., 2020) or triplet (Chen and Qian, 2020; Mao et al., 2021).

Recently, learning four elements simultaneously sparks new research interests. Two promising directions have been pointed out by researchers. Cai et al. (2021) propose a two-stage method by extract-

ing the aspect term and opinion term first. Then these items are utilized to classify aspect category and sentiment polarity. Another method is based on generation model (Zhang et al., 2021a). By paraphrasing the input sentence, the quadruplet can be extracted in an end-to-end manner. In this work, we follow the generative direction and consider the order-free property of the quadruplet. To the best of our knowledge, this work is the first to study ASQP from the order perspective.

6.2 Data Augmentation

Data augmentation has been widely adopted in both the language and vision fields. We formulate the input and output of a model as X and Y , respectively. Previous data augmentation can be divided into three types. The first type is augmenting the input X . For example, image flipping, rotation and scaling all change X to seek improvements (Shorten and Khoshgoftaar, 2019). In the text tasks, back translation (Sugiyama and Yoshinaga, 2019) can also generate pseudo pairs through augmenting X . The main idea is that changing X does not affect its ground-truth label Y . Secondly, both X and Y are augmented. A promising work is mixup (Zhang et al., 2018), which constructs virtual training examples base on the prior knowledge that linear interpolations of feature vectors should lead to linear interpolations of the associated targets. Despite it is intuitive, it has shown effectiveness in many tasks (Sun et al., 2020).

The third one is augmenting Y . One recent work proposes virtual sequence as the target-side data augmentation (Xie et al., 2022) for sequence-to-sequence learning. It deals with typical generation tasks, which are closely connected with the order of words. Different from it, we exploit the characteristic of the generative ASQP task. Order permutations still provide ground-truth labels. Then we think that different orders are just similar to seeing a picture from different perspectives, i.e. different views. Therefore, combining multiple template orders can prevent the model from being biased to superficial patterns, and help it to comprehensively understand the essence of the task.

7 Conclusion

In this work, we study aspect sentiment quad prediction (ASQP) from the template order perspective. We hypothesize that different orders provide various views of the quadruplet. In light of this hy-

pothesis, a simple but effective method is proposed to identify the most proper orders, and further combine multiple proper templates as data augmentation to improve the ASQP task. Specifically, we use the pre-trained language model to select the orders with minimal entropy. By fine-tuning the pre-trained model with these template orders, our model achieves state-of-the-art performances.

Limitations

Our work is the first attempt to improve the ASQP task by combining multiple template orders as data augmentation. Despite state-of-the-art performance, our work still have limitations which may guide the direction of future work.

Firstly, we use the entropy to select the proper template orders. The smaller entropy value indicates that the target sequence is better fitting with the pre-trained language model. However, there may be other criteria for template order selection which can better fine-tune the pre-trained language model to support the ASQP task.

Secondly, in the experiment, we simply select the top- k template orders for data augmentation. This can be treated as a greedy strategy for the combination. However, each of the top- k orders may not supplement well to each other. More advanced strategies may be designed to select template orders for data augmentation.

Thirdly, we only consider augmenting the target sequences in the model training, while augmenting both the input and out sequences may bring more performance improvement.

Acknowledgements

We sincerely thank all the anonymous reviewers for providing valuable feedback. This work is supported by the key program of the National Science Fund of Tianjin, China (Grant No. 21JCZDJC00130), the Basic Scientific Research Fund, China (Grant No. 63221028), the National Science and Technology Key Project, China (Grant No. 2021YFB0300104).

References

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. [ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction](#). In *Proceedings of the 2021 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 2069–2079.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. [Aspect-category based sentiment analysis with hierarchical graph convolutional network](#). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 833–843.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. [Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 340–350.

Zhuang Chen and Tiejun 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 (ACL)*, pages 3685–3694.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. [Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction](#). In *Findings of the Association for Computational Linguistics (ACL)*, page 45–57.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [Simcse: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

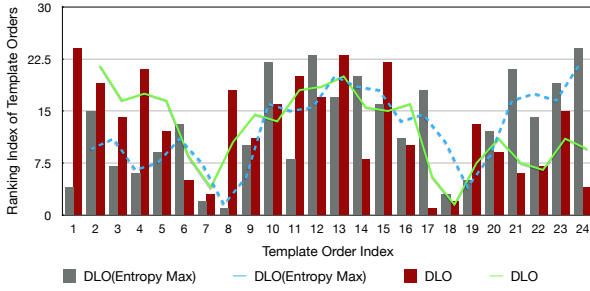
Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. [CAN: Constrained attention networks for multi-aspect sentiment analysis](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4601–4610.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. [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.

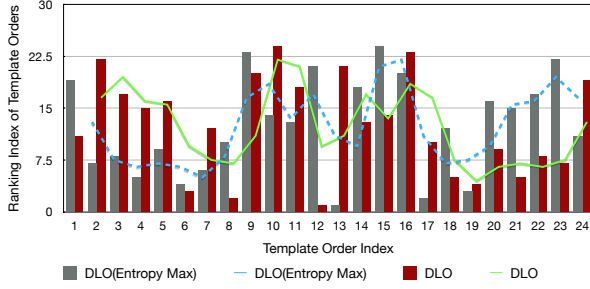
Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 8086–8098.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. [Exploring sequence-to-sequence learning in aspect term extraction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3538–3547.

- Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. [A joint training dual-mrc framework for aspect based sentiment analysis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 13543–13551.
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. [Structured prediction as translation between augmented natural languages](#). In *The Ninth International Conference on Learning Representations (ICLR)*, pages 1–26.
- Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. [Knowing what, how and why: A near complete solution for aspect-based sentiment analysis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 8600–8607.
- Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. [Semeval-2016 task 5: Aspect based sentiment analysis](#). In *International workshop on semantic evaluation*, pages 19–30.
- Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. [Semeval-2015 task 12: Aspect based sentiment analysis](#). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.
- Maria Pontiki, Dimitrios 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research (JMLR)*, 21(140):1–67.
- Connor Shorten and Taghi M Khoshgoftaar. 2019. [A survey on image data augmentation for deep learning](#). *Journal of big data*, 6(1):1–48.
- Amame Sugiyama and Naoki Yoshinaga. 2019. [Data augmentation using back-translation for context-aware neural machine translation](#). In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44.
- Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. [Mixup-transformer: Dynamic data augmentation for NLP tasks](#). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3436–3440.
- Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. [Order matters: Sequence to sequence for sets](#). In *The International Conference on Learning Representations (ICLR)*, pages 1–11.
- Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. [Target-aspect-sentiment joint detection for aspect-based sentiment analysis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 9122–9129.
- Ronald J Williams and David Zipser. 1989. [A learning algorithm for continually running fully recurrent neural networks](#). *Neural computation*, 1(2):270–280.
- Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, and Rui Yan. 2022. [Target-side input augmentation for sequence to sequence generation](#). In *International Conference on Learning Representations (ICLR)*, pages 1–18.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–15.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. [mixup: Beyond empirical risk minimization](#). In *International Conference on Learning Representations (ICLR)*, pages 1–13.
- Mi Zhang and Tiejun Qian. 2020. [Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549.
- Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. [Aspect sentiment quad prediction as paraphrase generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9209–9219.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. [Towards generative 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 (ACL-IJCNLP)*, pages 504–510.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. [A survey on aspect-based sentiment analysis: Tasks, methods, and challenges](#). *arXiv preprint arXiv:2203.01054*.
- He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. [SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3239–3248.



(a) Rest15



(b) Rest16

Figure 7: Ranking position of different template orders from DLO and DLO(Entropy Max).

A Appendix

A.1 Software and Hardware

We use Pytorch to implement all the models (Python 3.7). The operating system is Ubuntu 18.04.6. We use a single NVIDIA A6000 GPU with 48GB of RAM.

A.2 Full Pilot Experimental Results

The full experimental results of template permutations are presented in Table 7, Table 8, Table 9 and Table 10. The results of each table are sorted by F1 scores in an ascending order.

A.3 Results of DLO

Since the proposed DLO choose templates for the whole training set, we plot the ranking position of each template order in Figure 7. Here the horizon axis indicates that the template indexes which are ordered by F1 score from Table 9 and Table 10 in an ascending order. Then we can see that DLO can choose better-performed templates, where the ranking positions of $O_i, i \in [17, 24]$ are small. In contrary, DLO(Entropy Max) selects template orders that performed worse.

Template	Rest15		
	Pre	Rec	F1
$x_{sp}, x_{ac}, x_{at}, x_{ot}$	45.55	46.34	45.94
$x_{ot}, x_{at}, x_{ac}, x_{sp}$	45.43	47.02	46.21
$x_{at}, x_{ac}, x_{sp}, x_{ot}$	46.15	47.02	46.58
$x_{ot}, x_{ac}, x_{at}, x_{sp}$	46.04	47.25	46.63
$x_{sp}, x_{at}, x_{ac}, x_{ot}$	46.34	47.25	46.79
$x_{sp}, x_{ot}, x_{at}, x_{ac}$	46.12	47.52	46.81
$x_{ot}, x_{ac}, x_{sp}, x_{at}$	46.37	47.65	47.00
$x_{ot}, x_{sp}, x_{at}, x_{ac}$	46.37	47.90	47.12
$x_{at}, x_{sp}, x_{ac}, x_{ot}$	46.58	47.67	47.12
$x_{sp}, x_{at}, x_{ot}, x_{ac}$	46.44	47.90	47.15
$x_{at}, x_{sp}, x_{ot}, x_{ac}$	46.53	47.90	47.20
$x_{ac}, x_{sp}, x_{at}, x_{ot}$	46.67	47.85	47.25
$x_{at}, x_{ot}, x_{ac}, x_{sp}$	46.67	47.87	47.26
$x_{ot}, x_{sp}, x_{ac}, x_{at}$	46.78	47.90	47.33
$x_{at}, x_{ac}, x_{ot}, x_{sp}$	46.72	47.97	47.34
$x_{ac}, x_{at}, x_{ot}, x_{sp}$	46.75	48.10	47.41
$x_{ac}, x_{ot}, x_{at}, x_{sp}$	47.07	47.85	47.46
$x_{at}, x_{ot}, x_{sp}, x_{ac}$	46.91	48.23	47.56
$x_{ot}, x_{at}, x_{sp}, x_{ac}$	46.73	48.43	47.56
$x_{sp}, x_{ac}, x_{ot}, x_{at}$	47.29	48.03	47.66
$x_{sp}, x_{ot}, x_{ac}, x_{at}$	47.29	48.20	47.74
$x_{ac}, x_{at}, x_{sp}, x_{ot}$	47.29	48.28	47.78
$x_{ac}, x_{ot}, x_{sp}, x_{at}$	47.41	48.33	47.86
$x_{ac}, x_{sp}, x_{ot}, x_{at}$	47.60	48.75	48.17

Table 7: Evaluation results on Rest15, which are sorted by F1 scores.

Template	Rest16		
	Pre	Rec	F1
$x_{ot}, x_{at}, x_{sp}, x_{ac}$	56.04	58.17	57.09
$x_{ot}, x_{sp}, x_{at}, x_{ac}$	56.15	58.52	57.31
$x_{at}, x_{ac}, x_{sp}, x_{ot}$	56.71	58.52	57.60
$x_{ac}, x_{ot}, x_{at}, x_{sp}$	56.73	58.55	57.62
$x_{ot}, x_{ac}, x_{sp}, x_{at}$	56.78	59.00	57.87
$x_{ac}, x_{ot}, x_{sp}, x_{at}$	57.14	58.72	57.92
$x_{sp}, x_{ot}, x_{at}, x_{ac}$	56.87	59.07	57.95
$x_{ac}, x_{sp}, x_{ot}, x_{at}$	56.89	59.10	57.98
$x_{ot}, x_{at}, x_{ac}, x_{sp}$	56.95	59.07	57.99
$x_{ac}, x_{at}, x_{sp}, x_{ot}$	56.91	59.17	58.02
$x_{ot}, x_{ac}, x_{at}, x_{sp}$	57.14	58.97	58.04
$x_{sp}, x_{ac}, x_{at}, x_{ot}$	57.77	59.07	58.41
$x_{at}, x_{sp}, x_{ac}, x_{ot}$	57.35	59.60	58.45
$x_{at}, x_{ac}, x_{ot}, x_{sp}$	57.49	59.57	58.51
$x_{ot}, x_{sp}, x_{ac}, x_{at}$	57.58	59.77	58.66
$x_{sp}, x_{at}, x_{ac}, x_{ot}$	57.70	59.87	58.76
$x_{sp}, x_{ac}, x_{ot}, x_{at}$	57.71	59.87	58.77
$x_{at}, x_{ot}, x_{ac}, x_{sp}$	57.67	59.98	58.80
$x_{sp}, x_{ot}, x_{ac}, x_{at}$	58.14	59.82	58.97
$x_{at}, x_{ot}, x_{sp}, x_{ac}$	57.68	60.35	58.99
$x_{sp}, x_{at}, x_{ot}, x_{ac}$	57.93	60.13	59.01
$x_{ac}, x_{at}, x_{ot}, x_{sp}$	58.12	60.08	59.08
$x_{at}, x_{sp}, x_{ot}, x_{ac}$	57.95	60.33	59.11
$x_{ac}, x_{sp}, x_{at}, x_{ot}$	58.11	60.33	59.20

Table 8: Evaluation results on Rest16, which are sorted by F1 scores.

Template							Rest15		
							Pre	Rec	F1
[SP] x_{sp} [AT] x_{at} [AC] x_{ac} [OT] x_{ot}	45.34	46.39	45.86						
[SP] x_{sp} [AC] x_{ac} [OT] x_{ot} [AT] x_{at}	46.24	46.67	46.45						
[AC] x_{ac} [SP] x_{sp} [OT] x_{ot} [AT] x_{at}	46.28	47.37	46.81						
[SP] x_{sp} [OT] x_{ot} [AT] x_{at} [AC] x_{ac}	46.49	47.47	46.98						
[AT] x_{at} [SP] x_{sp} [AC] x_{ac} [OT] x_{ot}	46.40	47.85	47.11						
[OT] x_{ot} [SP] x_{sp} [AC] x_{ac} [AT] x_{at}	46.64	47.62	47.12						
[OT] x_{ot} [AT] x_{at} [SP] x_{sp} [AC] x_{ac}	46.97	47.75	47.35						
[AC] x_{ac} [AT] x_{at} [SP] x_{sp} [OT] x_{ot}	46.90	47.97	47.43						
[AT] x_{at} [AC] x_{ac} [SP] x_{sp} [OT] x_{ot}	47.00	48.08	47.53						
[AC] x_{ac} [OT] x_{ot} [SP] x_{sp} [AT] x_{at}	47.14	48.23	47.67						
[SP] x_{sp} [AT] x_{at} [OT] x_{ot} [AC] x_{ac}	47.06	48.40	47.72						
[AC] x_{ac} [SP] x_{sp} [AT] x_{at} [OT] x_{ot}	47.13	48.40	47.76						
[SP] x_{sp} [AC] x_{ac} [AT] x_{at} [OT] x_{ot}	47.27	48.35	47.80						
[OT] x_{ot} [SP] x_{sp} [AT] x_{at} [AC] x_{ac}	47.51	48.18	47.84						
[SP] x_{sp} [OT] x_{ot} [AC] x_{ac} [AT] x_{at}	47.53	48.18	47.85						
[AT] x_{at} [AC] x_{ac} [OT] x_{ot} [SP] x_{sp}	47.77	48.30	48.03						
[OT] x_{ot} [AT] x_{at} [AC] x_{ac} [SP] x_{sp}	47.57	48.55	48.06						
[AT] x_{at} [OT] x_{ot} [SP] x_{sp} [AC] x_{ac}	47.46	48.75	48.10						
[AC] x_{ac} [OT] x_{ot} [AT] x_{at} [SP] x_{sp}	48.14	48.20	48.17						
[OT] x_{ot} [AC] x_{ac} [SP] x_{sp} [AT] x_{at}	47.93	48.75	48.33						
[AT] x_{at} [SP] x_{sp} [OT] x_{ot} [AC] x_{ac}	48.36	49.08	48.72						
[OT] x_{ot} [AC] x_{ac} [AT] x_{at} [SP] x_{sp}	48.55	48.91	48.73						
[AC] x_{ac} [AT] x_{at} [OT] x_{ot} [SP] x_{sp}	48.58	49.01	48.79						
[AT] x_{at} [OT] x_{ot} [AC] x_{ac} [SP] x_{sp}	48.46	49.46	48.95						

Table 9: Evaluation results on Rest15, which are sorted by F1 scores.

Template							Rest16		
							Pre	Rec	F1
[AT] x_{at} [SP] x_{sp} [AC] x_{ac} [OT] x_{ot}	56.36	58.80	57.55						
[SP] x_{sp} [AC] x_{ac} [OT] x_{ot} [AT] x_{at}	57.17	58.97	58.06						
[AC] x_{ac} [SP] x_{sp} [AT] x_{at} [OT] x_{ot}	57.39	58.80	58.08						
[AC] x_{ac} [SP] x_{sp} [OT] x_{ot} [AT] x_{at}	57.07	59.17	58.10						
[AC] x_{ac} [AT] x_{at} [OT] x_{ot} [SP] x_{sp}	57.20	59.32	58.24						
[OT] x_{ot} [SP] x_{sp} [AC] x_{ac} [AT] x_{at}	57.31	59.25	58.26						
[AT] x_{at} [AC] x_{ac} [SP] x_{sp} [OT] x_{ot}	57.30	59.35	58.31						
[OT] x_{ot} [AT] x_{at} [AC] x_{ac} [SP] x_{sp}	57.29	59.42	58.34						
[SP] x_{sp} [OT] x_{ot} [AC] x_{ac} [AT] x_{at}	57.40	59.52	58.44						
[SP] x_{sp} [AT] x_{at} [AC] x_{ac} [OT] x_{ot}	57.77	59.17	58.46						
[OT] x_{ot} [AT] x_{at} [SP] x_{sp} [AC] x_{ac}	57.63	59.82	58.70						
[AC] x_{ac} [AT] x_{at} [SP] x_{sp} [OT] x_{ot}	57.64	59.80	58.70						
[SP] x_{sp} [OT] x_{ot} [AT] x_{at} [AC] x_{ac}	57.81	59.95	58.86						
[AC] x_{ac} [OT] x_{ot} [AT] x_{at} [SP] x_{sp}	58.04	59.82	58.92						
[AC] x_{ac} [OT] x_{ot} [SP] x_{sp} [AT] x_{at}	57.87	60.02	58.93						
[SP] x_{sp} [AC] x_{ac} [AT] x_{at} [OT] x_{ot}	58.29	59.80	59.03						
[AT] x_{at} [AC] x_{ac} [OT] x_{ot} [SP] x_{sp}	58.11	60.02	59.05						
[OT] x_{ot} [AC] x_{ac} [AT] x_{at} [SP] x_{sp}	58.13	60.10	59.10						
[AT] x_{at} [OT] x_{ot} [SP] x_{sp} [AC] x_{ac}	58.25	60.05	59.14						
[OT] x_{ot} [AC] x_{ac} [SP] x_{sp} [AT] x_{at}	57.64	60.85	59.20						
[AT] x_{at} [OT] x_{ot} [AC] x_{ac} [SP] x_{sp}	58.27	60.33	59.28						
[AT] x_{at} [SP] x_{sp} [OT] x_{ot} [AC] x_{ac}	58.03	60.63	59.30						
[OT] x_{ot} [SP] x_{sp} [AT] x_{at} [AC] x_{ac}	58.72	60.45	59.57						
[SP] x_{sp} [AT] x_{at} [OT] x_{ot} [AC] x_{ac}	58.50	60.75	59.60						

Table 10: Evaluation results on Rest16, which are sorted by F1 scores.