# Open-Domain Aspect-Opinion Co-Mining with Double-Layer Span Extraction

Mohna Chakraborty*
Iowa State University
Ames, Iowa, USA

Adithya Kulkarni*
Iowa State University
Ames, Iowa, USA

Qi Li
Iowa State University
Ames, Iowa, USA

## ABSTRACT

The aspect-opinion extraction tasks extract aspect terms and opinion terms from reviews. The supervised extraction methods achieve state-of-the-art performance but require large-scale human-annotated training data. Thus, they are restricted for open-domain tasks due to the lack of training data. This work addresses this challenge and simultaneously mines aspect terms, opinion terms, and their correspondence in a joint model. We propose an Open-Domain Aspect-Opinion Co-Mining (ODAO) method with a Double-Layer span extraction framework. Instead of acquiring human annotations, ODAO first generates weak labels for unannotated corpus by employing rules-based on universal dependency parsing. Then, ODAO utilizes this weak supervision to train a double-layer span extraction framework to extract aspect terms (ATE), opinion terms (OTE), and aspect-opinion pairs (AOPE). ODAO applies canonical correlation analysis as an early stopping indicator to avoid the model over-fitting to the noise to tackle the noisy weak supervision. ODAO applies a self-training process to gradually enrich the training data to tackle the weak supervision bias issue. We conduct extensive experiments and demonstrate the power of the proposed ODAO. The results on four benchmark datasets for aspect-opinion co-extraction and pair extraction tasks show that ODAO can achieve competitive or even better performance compared with the state-of-the-art fully supervised methods.

## CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Machine learning*.

## KEYWORDS

Review Analysis, Data Mining, Natural Language Processing.

**ACM Reference Format:**
Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2022. Open-Domain Aspect-Opinion Co-Mining with Double-Layer Span Extraction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539386

*Both authors contributed equally to this paper.



*KDD '22, August 14–18, 2022, Washington, DC, USA*

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539386

## 1 INTRODUCTION

Understanding customer requirements is crucial for business development. Due to the massive volume of reviews, many businesses need to conduct cost-effective review analysis to enhance their services. Review analysis consists of multiple tasks including aspect term extraction (*ATE*), opinion term extraction (*OTE*), aspect-opinion pair extraction (*AOPE*), aspect-based sentiment analysis (*ABSA*), aspect-specified opinion extraction (*ASOE*), etc. In review analysis, aspect terms describe the product or service attributes, and opinion terms describe the reviewer's opinion towards the corresponding product or aspects of a product. Considering the review *"the wine list is extensive and impressive ."*, the aspect term is *"wine list"*, the corresponding opinion terms are *"extensive"* and *"impressive"*, and aspect-opinion pairs are *("wine list","extensive")* and *("wine list","impressive")*. Our work aims to simultaneously mine aspect terms, opinion terms, and their correspondence.

Early works focusing on *ATE*, *OTE*, and *AOPE* tasks [11, 23, 24] are rule-based methods that utilize features such as corpus-level statistics and dependency parse trees. Frequent patterns are mined first and then used to form rules. These rules can work on various domains of reviews. However, high-quality rules can be sparse and of low coverage due to the variation of language expressions, and some low-quality rules may introduce noise in the results. These rules also face challenges for complex aspect-opinion expressions (for instance, one aspect may correspond to many opinion terms).

Existing works on *ATE* [14, 18, 31–33, 36], *OTE* [30, 32, 35, 39], and *AOPE* [2, 8, 12] tasks achieved state-of-the-art results using deep neural networks trained on human-annotated labels. These supervised methods can learn the complex relationships between aspect terms and opinion terms. However, these methods rely on human-annotated datasets, which can be expensive to obtain. Due to their dependency on the labeled dataset, these methods may perform poorly in the resource-scarce domains.

Several semi-supervised methods are proposed to tackle the issue of insufficient labeled data. Similar to rule-based methods, semi-supervised methods [4, 39] also mine rules. These methods utilize a human-annotated dataset to mine rules of high quality. These mined rules are then utilized to annotate unlabeled corpora. The weakly labeled and human-annotated datasets are used to train deep neural networks. These methods improve the performance for cross-domain tasks but still require a related corpus with human annotations.

Our work[1] aims to develop a framework for open-domain aspect-opinion co-mining tasks with no human-annotated corpus. We adopt the findings of previous rule-based methods [24] to form some high-quality rules that apply to a wide range of domains.

[1]The code can be found at https://github.com/kulkarniadithya/ODAO

These modified rules are then applied to annotate review corpora. Compared with human-annotated labels, the weak labels provided by the rules are biased and noisy. To handle these problems, we propose a novel double-layer span extraction model ODAO.

The proposed ODAO simultaneously conducts three tasks, namely, *ATE*, *OTE*, and *AOPE*. We further decouple the task of *AOPE* into two sub-tasks, aspect specified opinion extraction (*ASOE*) and opinion specified aspect extraction (*OSAE*). The four tasks, *ATE*, *OTE*, *ASOE*, and *OSAE*, are closely related and can mutually enhance each other. Among the four tasks, *ATE* and *OSAE* tasks have similar goals to extract aspect terms, and *OTE* and *ASOE* tasks have similar goals to extract opinion terms. Further, *ASOE* and *OSAE* can be considered the subsequence tasks of *ATE* and *OTE*. To jointly model the four tasks in one framework, we propose a double-layer architecture with a BERT-based span extractor for each task.

We further utilize the correlation among the tasks to tackle the problem of bias and noise in the weakly labeled training data. Previous work [15] notices that early stopping can prevent the model from over-fitting to the noisy annotated labels. However, when to stop remains a challenge in the absence of ground truth labels. We use the following observations to tackle this challenge. Intuitively, the tasks with the same goal should agree on their interpretations for the same reviews. For example, aspect terms extracted by the *OSAE* module should also be extracted by the *ATE* module. Therefore, the correlations between the coupled tasks can indicate the learning status. When the hidden representations of the coupled module are maximally correlated, it implies that the coupled tasks are properly trained. Therefore, we adopt the canonical correlation analysis (*CCA*) [1, 10, 13] on the hidden representations of the reviews to measure this correlation and use *CCA* as early stopping criteria during training to avoid the model over-fitting to the noisy and biased labels. Furthermore, if an unlabeled review receives agreed predictions of all four sub-tasks, then the review is likely to be predicted correctly. Thus ODAO adopts a self-training idea, adding such highly confident reviews with their predicted labels to the training pool to enrich the training data and then re-training the model.

We conduct extensive experiments on various benchmark datasets from different domains and evaluate the aspect terms extracted by the *ATE* module, opinion terms extracted by the *OTE* module, and the aspect-opinion pairs extracted by the combination of *ASOE* and *OSAE* modules. The experimental results show that ODAO outperforms previous semi-supervised methods and achieves competitive performance to the state-of-the-art fully supervised methods for all the three tasks of *ATE*, *OTE*, and *AOPE*, even though ODAO uses only a small amount of rules to obtain weakly labeled training data. The experimental results demonstrate the effectiveness of the proposed ODAO in real applications.

In summary, the main contributions of this paper are as follows:

- We propose ODAO to simultaneously extract aspect terms, opinion terms, and aspect-opinion pairs in a review. To the best of our knowledge, this is the first work that conducts these tasks for open-domain review analysis with weak supervision.
- We design a double-layer span extraction framework to jointly model the tasks from different aspects. Specifically, ODAO jointly models *ATE*, *OTE*, *ASOE*, and *OSAE* tasks and fully consider their correlation.
- The proposed ODAO is resilient to biased, noisy training data provided by rules. Specifically, *CCA* as early stopping criteria prevents the model from over-fitting to the noisy labels, and the self-training process enriches training data to address the training bias problem.
- Extensive experiments on benchmark datasets from various domains validate the effectiveness of the proposed ODAO.

## 2 RELATED WORKS

Review analysis, which aims to analyze people's detailed insights towards a product or service, has become an extensive research topic in natural language processing. There are many sub-tasks in this domain. For example, aspect term extraction (*ATE*) [29] aims to extract aspect terms in the reviews, aspect-opinion co-extraction [35] aims to extract aspect and opinion terms, aspect-opinion pair extraction (*AOPE*) [8] aims to extract aspect terms and their corresponding opinion terms, and aspect-based sentiment analysis (*ABSA*) [27] aims to extract aspect terms and classify their corresponding sentiments . This work focuses on aspect term and opinion term co-extraction and pair-wise extraction. These are considered fundamental sub-tasks of review analysis that have gained much attention over recent years.

**Aspect Term Extraction (*ATE*).** Unsupervised approaches designed for open-domain *ATE* task mainly include methods based on frequent pattern mining [11], topic modeling [19], and neural networks [9, 16]. However, there is a significant gap in terms of performance compared with supervised models trained on deep neural networks [29, 36, 37]. These works extract aspect terms from reviews without considering the information of the opinion terms.

**Aspect and Opinion Term Co-Extraction**. Supervised and semi-supervised models have been proposed for this task [4, 24, 28, 39]. Traditional methods [24, 28, 42] treat co-extraction of the aspect and opinion terms in a pipeline-based manner using dependency parsing results. Deep neural network-based frameworks dominate the supervised models. Various methods are proposed to jointly extract aspect terms and opinion terms from reviews, such as by considering manual features [32], sharing information via attention mechanisms [33], and conducting transfer-learning for cross-domain aspect, and opinion terms co-extraction [30]. These approaches have outperformed the *ATE* methods thanks to considering opinion terms and interactions among opinion terms and aspect terms; however, none of these works considered the aspect terms and opinion terms as pairs.

**Aspect-Opinion Pair Extraction (*AOPE*).** There are several strategies for the *AOPE* task. One strategy is to extract aspect terms first and then extract the aspect-oriented opinion terms. For example, Gao et al. [8] utilizes a span-based extraction mechanism to first extract aspect term spans and then use the aspect term span along with the reviews in a question-answering fashion to extract the opinion terms. Another strategy is to extract aspect and opinion terms and score the relation of each pair jointly, or in a pipeline, fashion [2, 40]. For example, Zhao et al. [40] proposes a multi-task

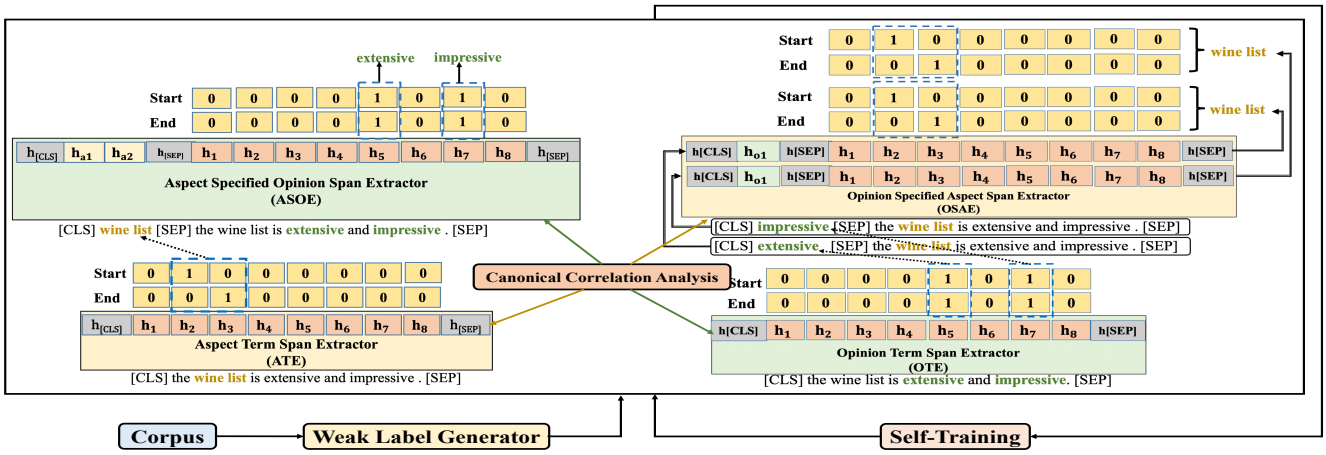



**Figure 1: ODAO architecture**

**Table 1: Summary of Notations**

| Notation | Definition |
|---|---|
| $\mathcal{D}$ | the review corpus |
| $\mathcal{D}'_{labeled}$ | the weak labeled train set of the corpus |
| $\mathcal{D}'_{unlabeled}$ | the unlabeled train set of the corpus |
| $R$ | a review comprised of $k$ tokens |
| $A$ | a set of aspect terms in each review |
| $O$ | a set of opinion terms in each review |
| $P$ | a set of aspect-opinion pairs in each review |

learning framework to jointly learn the span boundaries and span relations. In our work, we adopt a fusion of these two strategies and design a double-layer span extraction framework.

The state-of-the-art performances are achieved using deep neural networks, which rely on large-scale human-annotated training data. These deep neural models may face significant challenges when training data is insufficient. In this regard, some methods [4, 29] propose to add additional training data with pseudo-labels. For example, Dai and Song [4] proposes to mine rules from labeled data and use those rules to generate pseudo-labels on auxiliary datasets to enlarge the training data. However, these methods still require human-annotated data. For open-domain tasks with insufficient or no annotated data, there are some unsupervised [16] and semi-supervised [39] methods. However, there is a big gap in the performance compared to supervised methods.

In this work, we adopt the usage of weakly labeled data [3] and propose a double-layer architecture with special mechanisms designed for the noise and bias issues of the weak supervision. Without resorting to intensive human effort to label the training data, the proposed ODAO achieve comparable performance with the prior supervised works.

# 3 METHODOLOGY

In this section, we first provide an overview of the problem and the proposed framework in Section 3.1 and Section 3.2, respectively. Then, the weak label generation is discussed in Section 3.3, along with a detailed description of the model in Section 3.4.

## 3.1 Problem Formulation

Suppose there is a corpus $\mathcal{D}$ that contains $N$ unlabeled reviews. For a review $R = \{w_1, w_2, ..., w_k\}$ consisting of $k$ tokens, the task of aspect term extraction (ATE) is to extract all the spans of aspect terms ($A = \{A_1, A_2, .., A_i\}$), opinion term extraction (OTE) is to extract all the spans of opinion terms ($O = \{O_1, O_2, ..., O_j\}$), and aspect-opinion pair extraction (AOPE) is to extract all aspect-opinion pairs ($P = \{(A_i, O_j), ..\}$) from the review $R$. Some frequently used notations are summarized in Table 1.

## 3.2 Overview

Figure 1 provides an illustration of the proposed model. Given a corpus $\mathcal{D}$, we obtain weak labels employing the weak label generator. The weak label generator uses four rules that consider the correlation between opinion terms and aspect terms to extract opinion terms ($O$) and corresponding aspect terms ($A$) in the review. The weakly labeled reviews are added to the set $\mathcal{D}'_{labeled}$ while the unlabeled reviews are added to the set $\mathcal{D}'_{unlabeled}$. Detailed discussion for weak label generator is provided in Section 3.3. The weakly labeled train set ($\mathcal{D}'_{labeled}$) is then utilized for training a double-layer span extraction model for *ATE*, *OTE*, *ASOE*, and *OSAE* tasks. To avoid over-fitting to the noise in the weakly labeled train set, canonical correlation analysis (*CCA*) between the *ATE*, *OSAE*, and *OTE*, *ASOE* modules are used as early stopping criteria. Since $\mathcal{D}'_{labeled}$ only covers a small portion of the corpus $\mathcal{D}$, we employ a self-training strategy to enrich the training data and improve model performance. At the end of each iteration, the trained model predicts on the unlabeled set $\mathcal{D}'_{unlabeled}$ and predictions with high confidence are adopted as pseudo labels. This pseudo labeled data is added to $\mathcal{D}'_{labeled}$ to re-train the model in the next iteration. After the training iterations, the trained model can then be used on the test set $\mathcal{D}_{test}$ of the corpus for the tasks of *ATE*, *OTE*, and *AOPE*. Detailed discussion is provided in Section 3.4.

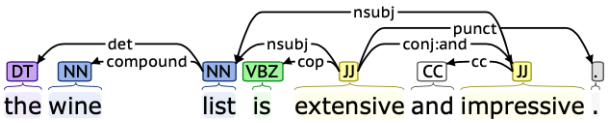



**Figure 2: An example of Dependency parse tree**

## 3.3 Weak Label Generation

We briefly introduce how we generate weak labels for open-domain aspect-opinion co-mining tasks to establish training data without human annotations. Please refer to the Appendix for more technical details. Previous work [24] handcrafts rules based on dependency parse trees under the assumption that aspect terms are generally nouns and opinion terms are generally adjectives. We adopt a simple rule $AspectTerm = NN \leftarrow nsubj \leftarrow JJ(root) = OpinionTerm$. This rule states that if a noun word ($NN$) is the "nsubj" of an adjective word ($JJ$) and the adjective word is the root of this review, then the noun word is an aspect term and the adjective word is an opinion term. An example review with its dependency parse tree[2] is illustrated in Figure 2. According to this rule, "list" is an aspect term, and "extensive" is the corresponding opinion term. Similarly, for the same aspect term "list", "impressive" is another opinion term.

To further improve the quality of this rule, we extend it with some additional considerations. Specifically, this rule may provide incorrect span boundaries for aspect terms and opinion terms. For example, the aspect term should be "wine list" instead of "list". To handle this problem, we combine noun words with the *compound* relations as they are likely to form a phrase. We also obtain high confidence phrases in the corpus employing phrase mining tool [25] and use this information to correct the span boundary of extracted aspect terms and opinion terms. We extend this rule to handle $conj$ relations to detect more aspect and opinion terms. For more technical details, please refer to Appendix.

With the updated rules, the example review in Figure 2 will be labeled with "wine list" as the aspect term, "extensive" and "impressive" as the opinion terms, and ("wine list", "extensive") and ("wine list", "impressive") as the aspect-opinion pair.

Only the reviews that can strictly follow the rules are labeled and added to $\mathcal{D}'_{labeled}$. Consequently, the reviews in $\mathcal{D}'_{labeled}$ biasedly represent the whole corpus. The rules can only label a small portion of the corpus (25.68% of the SemEval 14 restaurant dataset) but can achieve relatively high precision. With the weakly labeled training data characteristics, we propose ODAO in the following section.

## 3.4 Model Description

*3.4.1* ***Encoder***. The role of the encoder in our model is to provide rich semantic, syntactic, and context-sensitive information for each token in the input review. In this work, we use pre-trained Bidirectional Encoder Representations from Transformers (BERT) [6] as base encoder for our model. BERT provides rich contextualized word representations, and its bidirectional self-attention unifies the features of self-attention and cross-attention. Motivated by [41], our framework is build upon four independent encoders to tackle the tasks of *ATE*, *OTE*, *OSAE*, and *ASOE*.

[2]We use CORENLP dependency parser for the purpose

*3.4.2* ***Aspect/Opinion Term Extractor***. Most of the previous works on the *ATE* task approach it as a sequential labeling problem based on the BMES [38], or BIO [36] tagging schemes, similar to named entity recognition (NER) tasks. Recent studies in weakly supervised NER tasks find that the sequential labeling schemes do not work well with noisy labels, especially noisy boundaries. Motivated by advances in relation extraction [34], we adopt a span-based instead of a BIO or BMES tagging scheme for the tasks.

For simplicity, we describe the aspect term span extractor (*ATE*) module in this section. The opinion term span extractor (*OTE*) module has a similar framework. Given an input review $R = \{w_1, w_2, ..., w_k\}$ consisting of $k$ tokens, $[CLS]$ and $[SEP]$ tokens are appended at the start and end of the review, respectively. BERT encoder is then used to encode the review to obtain hidden representations $H = \{h_{[CLS]}, h_1, h_2, ..., h_{[SEP]}\}$, where the embedding dimension of each $h_i$ is $d_h$ and $|H| = N'$. These hidden representations are passed to a linear layer that applies linear transformation on the hidden representations to provide score for start span ($h_{i_s}$) and end span ($h_{i_e}$).

$$\begin{aligned} y_i &= h_i * \mathcal{W}^T + b, \\ h_{i_s} &= y_i[0], \\ h_{i_e} &= y_i[1], \end{aligned} \tag{1}$$

where $y_i \in \mathbb{R}^2$, $\mathcal{W} \in \mathbb{R}^{2*d_h}$, and $b \in \mathbb{R}^2$, respectively, and 2 represents the start and end span. $\mathcal{W}$ and $b$ are initialized randomly from $\mathcal{U}(-\sqrt{f}, \sqrt{f})$, where $f = \frac{1}{d_h}$.

The prediction is obtained from the scores of start span ($h_{i_s}$) and end span ($h_{i_e}$) as follows:

$$\begin{aligned} \hat{y}_i^s &= \begin{cases} 1, & \text{if } h_{i_s} > 0; \\ 0, & else. \end{cases} \\ \hat{y}_i^e &= \begin{cases} 1, & \text{if } h_{i_e} > 0; \\ 0, & else. \end{cases} \end{aligned} \tag{2}$$

Note that we obtain the set of aspect term spans in the review $R$ by matching each $\hat{y}_i^s = 1$ with its nearest $\hat{y}_j^e = 1$ such that $j \geq i$. The loss function is the averaged binary cross-entropy loss (BCE) between the predicted and labeled spans.

$$\mathcal{L}_{ATE} = \frac{\mathcal{L}_{ATE}^s + \mathcal{L}_{ATE}^e}{2} = \frac{\sum_{i=1}^{N'} \sum_{sp \in \{s,e\}} BCE(\hat{y}_i^{sp}, y_i^{sp})}{2}. \tag{3}$$

*3.4.3* ***Aspect Opinion Pair Extractor***. Aspect opinion pair extractor (*AOPE*) has more complexities than *ATE* and *OTE* tasks because of the complex pairing scenarios between the aspect terms and the opinion terms. One aspect term can pair with one opinion term (e.g. *"the wine list is extensive"*), multiple opinion terms (e.g. *"the wine list is extensive and impressive"*), and no opinion term (e.g. *"the wine list was given"*). Similarly, one opinion term can pair with one aspect term (e.g. *"the wine list is extensive"*), multiple aspect terms (e.g. *"the wine list and beer list are extensive"*), and no aspect term (e.g. *"it is extensive"*).

To model all pairing scenarios between the aspect terms and the opinion terms, we decouple the task of *AOPE* into

two sub-tasks, *ASOE* (aspect specified opinion extraction) and *OSAE* (opinion specified aspect extraction). For simplicity, we describe the *ASOE* module in this section. The *OSAE* module has a similar framework. Like the *ATE* module, we adopt a span-based scheme for the *ASOE* module. Given an input review $R = \{w_1, w_2, ..., w_k\}$ consisting of $k$ tokens and $A_p = \{AT_1, AT_2, ....\}$ as aspect term predictions from *ATE* module, each of the predicted span $AT_i = \{a_1, .., a_q\}$ is concatenated with the review $R$ as $I = \{[CLS], a_1, ..a_q, [SEP], w_1, w_2, ..., w_k, [SEP]\}$. If the *ATE* module predicts no aspect terms, then $A_p = \phi$ and $I = \{[CLS], [SEP], w_1, w_2, ..., w_k, [SEP]\}$.

BERT encoder is used to encode the review to obtain hidden representations $H = \{h_{[CLS]}, h_{a_1}, h_{a_2}, ..., h_{[SEP]}, h_1, ..., h_{[SEP]}\}$, where the embedding dimension of each $h_i$ is $d_h$ and $|H| = N'$. Similar to *ATE* module, a linear layer applies linear transformation on the hidden representation to provide score for start span ($h_{i_s}$) and end span ($h_{i_e}$) which is then utilized to obtain predictions $\hat{y}_i^s$ and $\hat{y}_i^e$. The set of opinion term spans are obtained by matching each $\hat{y}_i^s = 1$ with its nearest $\hat{y}_j^e = 1$ such that $j \geq i$. The loss function is also the averaged binary cross-entropy loss (BCE) between the predicted spans and labeled spans.

$$\mathcal{L}_{ASOE} = \frac{\mathcal{L}^s_{ASOE} + \mathcal{L}^e_{ASOE}}{2} = \frac{\sum_{i=1}^{N'} \sum_{sp \in \{s,e\}} BCE(\hat{y}_i^{sp}, y_i^{sp})}{2}. \tag{4}$$

During the testing phase, the predictions of *ASOE* and *OSAE* are aggregated to conduct the *AOPE* task. First, the prediction pairs with either aspect term or opinion term as *null* are removed. Among the remaining predictions, if the predictions of *ASOE* and *OSAE* modules match, the pair is considered as aspect-opinion pair; otherwise, it is discarded.

*3.4.4* ***Loss Function***. Finally, the overall loss for the model is the sum of loss from *ATE*, *OTE*, *ASOE*, and *OSAE* modules.

$$\mathcal{L} = \mathcal{L}_{ATE} + \mathcal{L}_{OTE} + \mathcal{L}_{ASOE} + \mathcal{L}_{OSAE}. \tag{5}$$

*3.4.5* ***Early Stopping***. The weakly labeled training data $\mathcal{D}'_{labeled}$ is biased and noisy. Motivated by [15], we employ early stopping to prevent the model from over-fitting to the label noise. Early stopping helps regularize model training and improve model generalization ability to unseen data. However, it is still challenging to decide on the stopping criteria due to the absence of ground truth labels. Previous work [15] uses a hyperparameter to pre-define the early stopping time, but different tasks may need different parameters, and it is hard to decide manually. It is also hard to tune without ground truth labels. To tackle this challenge, we propose to learn the proper early stopping criteria based on the weakly labeled training data.

Our intuition is that the modules with similar goals should interpret the same review similarly. If the model under-fits, the modules for different tasks interpret the review more independently, and thus the correlation is low. If the model over-fits, the modules will interpret the reviews from the perspective of their specific tasks, and thus the correlation is also low. Specifically, the hidden representations of the review can reflect such interpretation from the paired modules (i.e., *ATE* and *OSAE*, *OTE* and *ASOE*), and when the correlation of the hidden representations is maximized, it indicates that the model is properly trained. To measure this correlation, we adopt the Canonical correlation analysis (*CCA*) [1, 10, 13].

Let $H_{ATE}$ and $H_{OSAE}$ be the hidden representations of a review $R = \{w_1, w_2, ...w_k\}$ obtained by the encoders of *ATE* and *OSAE* modules, respectively. *CCA* seeks vectors $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^k$ such that the random variables $u^\intercal H_{ATE}$ and $v^\intercal H_{OSAE}$ maximize the correlation $\rho_1 = corr(u^\intercal H_{ATE}, v^\intercal H_{OSAE})$.

$$(u', v') = \underset{u,v}{\operatorname{argmax}}\ corr(u^\intercal H_{ATE}, v^\intercal H_{OSAE}), \tag{6}$$

let $\sum_{AO}$ be the cross-covariance matrix and $\sum_{AA}$, and $\sum_{OO}$ be co-variance matrices of $H_{ATE}$ and $H_{OSAE}$, respectively, then the function to maximize is

$$\rho_1 = \frac{u^\intercal \sum_{AO} v}{\sqrt{(u^\intercal \sum_{AA} u)(v^\intercal \sum_{OO} v)}}. \tag{7}$$

The estimation of co-variance matrices ($\sum_{AA}$, $\sum_{OO}$) with regularization helps detect over-fitting in the hidden representation ($H_{ATE}, H_{OSAE}$) [5]. Similarly, we can calculate the correlation score between *OTE* and *ASOE* modules for the same review as $\rho_2$.

For each epoch of training, the correlation score of the model is defined as:

$$\rho = \frac{\sum_M (\rho_1 + \rho_2)}{M}, \tag{8}$$

where $M$ refers to the number of reviews in the weakly labeled train set $\mathcal{D}'_{labeled}$. The correlation score of the model is essentially the average of the correlation scores over all reviews in $\mathcal{D}'_{labeled}$.

The correlation score is maximized when the hidden representations of the coupled tasks are maximally correlated. The model is properly trained at this stage and should stop training. Practically, the checkpoint with the maximum $\rho$ will be used as the final model.

*3.4.6* ***Self-Training***. The weakly labeled train set $\mathcal{D}'_{labeled}$ constrains the proposed model performance due to the low coverage of the weak label generator rules. Furthermore, the bias in $\mathcal{D}'_{labeled}$ can also influence model training. Self-training is adopted to enrich the training data and reduce bias. To control the noise level of the training data, we propose to select reviews based on the prediction confidence.

Given the model trained on $\mathcal{D}'_{labeled}$ with early stopping, it can be used to predict for the unlabeled reviews in $\mathcal{D}'_{unlabeled}$. We measure the prediction confidence based on the agreement level among the tasks on the predicted labels for the review. Specifically, we compute label agreement between *ATE*, *OSAE* modules, and between *OTE*, *ASOE* modules.

Let $A'_{ATE}$ and $A'_{OSAE}$ be the set of aspect term predictions by *ATE*, *OSAE* modules, respectively, and $O'_{OTE}$ and $O'_{ASOE}$ be the set of opinion term predictions by *OTE*, *ASOE* modules, respectively, for the review $R$. We use the symmetric difference between the sets to compute the disagreement among the module prediction as follows:

$$\gamma_R = A'_{ATE} \Delta A'_{OSAE} + O'_{OTE} \Delta O'_{ASOE}, \tag{9}$$

where $A \Delta B = (A - B) \bigcup (B - A)$ is the symmetric difference of two sets. All the modules agree on the predictions for a review $R$ if $|\gamma_R| = 0$, and such reviews are considered to be correctly predicted.

**Table 2: Statistics of the Datasets**

| Datasets | $\mathbb{S}_{14}l$ | | $\mathbb{S}_{14}r$ | | $\mathbb{S}_{15}r$ | | $\mathbb{S}_{16}r$ | |
|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| #sentences | 3045 | 800 | 3041 | 800 | 1315 | 685 | 2000 | 676 |
| #aspects | 2359 | 653 | 3693 | 1134 | 1205 | 542 | 1757 | 622 |
| #opinions | 2500 | 677 | 3512 | 1014 | 1217 | 516 | 1381 | 475 |

**Table 3: Statistics of Fan et al. [7] datasets**

| Datasets | $\mathbb{S}_{14}l$ | | $\mathbb{S}_{14}r$ | | $\mathbb{S}_{15}r$ | | $\mathbb{S}_{16}r$ | |
|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| #sentences | 1158 | 343 | 1627 | 500 | 754 | 325 | 1079 | 329 |
| #pairs | 1634 | 482 | 2643 | 865 | 1076 | 436 | 1512 | 457 |

All the correctly predicted reviews are adopted as pseudo-labeled data. The pseudo-labeled data is added in $\mathcal{D}'_{labeled}$ to enrich the training data, which is used for model training in the next iteration. This iterative process continues until the count of pseudo-labeled reviews in a given iteration is below a threshold.

# 4 EXPERIMENTS

In this section, we evaluate the proposed ODAO model on several benchmark datasets from various domains.

## 4.1 Datasets

The performance of the proposed ODAO is evaluated on four widely used datasets obtained from SemEval 2014 Task 4 [22] (SemEval-2014 Laptop or $\mathbb{S}_{14}l$, SemEval-2014 Restaurant or $\mathbb{S}_{14}r$), SemEval 2015 Task 12 [21] (SemEval-2015 Restaurant or $\mathbb{S}_{15}r$), and SemEval 2016 Task 5 [20] (SemEval-2016 Restaurant or $\mathbb{S}_{16}r$). The dataset statistics are provided in Table 2. The SemEval challenge only provides aspect term annotations for these datasets, so only *ATE* task is evaluated on the original annotations. For evaluation of *OTE* tasks, we utilize the annotations provided by Wang et al. [32] for $\mathbb{S}_{14}l$ and $\mathbb{S}_{14}r$ datasets, Wang et al. [33] for $\mathbb{S}_{15}r$ dataset, and Wu et al. [35] for $\mathbb{S}_{16}r$ dataset, to align with the baselines. For aspect-opinion pair evaluation, we utilize the annotations provided by Fan et al. [7], which only include the reviews that contain aspect-opinion pairs. Table 3 shows the dataset statistics.

## 4.2 Evaluation Metrics

We follow the same evaluation metrics as previous works [8]. We use the $F_1$ score to evaluate the performance of our model and compare with the baselines for *ATE*, *OTE*, and *AOPE* tasks. For aspect-opinion pair extraction, a pair is considered correct if the aspect term and corresponding opinion term are predicted correctly.

## 4.3 Baseline Methods

We compare our proposed method ODAO with the state-of-the-art approaches for *ATE*, *OTE*, and *AOPE* subtasks. These approaches can be partitioned into three categories.

***Aspect Term Extraction only.*** The following baselines focus on the stand-alone aspect term extraction task.

**PSTD** [29]: PSTD uses progressive self-training to add more training data with psuedo labels from auxiliary data.

**ABAE** [9]: ABAE employs an attention-based model to conduct *ATE* task in an unsupervised fashion.

**LCC+GBC** [16]: LCC+GBC employs a neural model that couples global (on sentence level) and local context (conveyed by neighboring words) to conduct *ATE* task in an unsupervised way.

**AutoNER** [26]: AutoNER utilizes "tie-or-break" labeling schema to conduct *ATE* tasks with dictionaries of aspect terms.

***Aspect-Opinion Term Co-Extraction.*** The following baselines conduct co-extraction of aspect terms and opinion terms. They are compared for *ATE* and *OTE* tasks.

**RINANTE** [4]: RINANTE trains a neural model on the SemEval training data with additional rule-labeled auxiliary data.

**DeepLogic** [31]: DeepLogic integrates deep learning with logic rules.

**DeepWMaxSAT** [35]: DeepWMaxSAT is a deep neural network model with logical reasoning and structured learning.

**GMTCMLA** [39]: GMTCMLA utilizes a small portion of human-annotated training data ( 200 randomly chosen training samples) to train a multi-task learning framework by modeling syntactic constraints through global inference.

**DP** [24]: DP is a rule-based approach that uses an opinion lexicon to identify opinion terms. These identified opinion terms are then used to extract aspect and opinion terms through double propagation.

***Aspect-Opinion Pair Extraction.*** The following baselines conduct pair extraction of aspect terms and opinion terms. They are compared for the *AOPE* task.

**QDSL** [8]: QDSL is a Question-Driven Span Labeling model to extract all the aspect–opinion pairs from reviews.

**SDRN** [2]: SDRN utilizes a multi-task learning framework to extract opinion entities and relations simultaneously.

**SpanMlt** [40]: SpanMlt develops a multi-task learning framework to jointly extract terms and score their relations.

For all baseline methods, we report their results according to their original publications. In addition to the state-of-the-art baseline methods, we also include fully supervised ODAO trained with the human-annotated training data, **FS-ODAO**. In FS-ODAO, we remove the early stop and self-training steps. FS-ODAO is trained in the same setting with other supervised baseline methods.

## 4.4 ODAO Setups

For each of the modules in ODAO, we choose the pre-trained uncased BERT ($BERT_{BASE}$) encoder with 12 attention heads, 12 hidden layers, and the hidden size of 768, resulting in 110M pre-trained parameters. The implementation is done in PyTorch, and we append a linear layer [3] on top of the BERT encoder for getting the scores for start and end spans. During the training process, we employ AdamW [17] to optimize the model parameters. The learning rate is set to $1e-5$, the batch size is set to 16. All the experiments are executed on one Nvidia GeForce RTX GPU. For each iteration, the execution approximately takes 30 minutes.

For the laptop domain, we use the raw text of SemEval-2014 Laptop ($\mathbb{S}_{14}l$) as training corpus, and for the restaurant domain, we combine the raw texts of SemEval-2014 Restaurant ($\mathbb{S}_{14}r$) and SemEval-2016 Restaurant ($\mathbb{S}_{16}r$) as training corpus. Note that only

[3] https://pytorch.org/docs/stable/generated/torch.nn.Linear.html

**Table 4: Results of *ATE* task from *ATE* module on SemEval dataset. We report the span-level $F_1$ scores on the test sets. Results of the baselines are reported from their original papers. - refers to unpublished results as of the date of writing (Feb. 2022).**

| Method | Human Effort | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|---|
| RINANTE | Gold Annotation | 80.16 | 86.45 | 69.90 | - |
| QDSL | | 84.27 | 87.85 | 77.72 | 83.34 |
| PSTD | | **86.91** | 88.75 | 75.82 | 82.56 |
| DeepWMaxSat | | 81.33 | 85.33 | - | 73.67 |
| **FS-ODAO** | | 85.93 | **88.77** | **83.39** | **86.15** |
| ABAE | None | 32.9 | 40.2 | | |
| LCC+GBC | | 36.1 | 41.2 | | |
| GMTCMLA | Sample Annotation | 56.08 | 76.51 | 61.75 | - |
| AutoNER | Dictionary | 65.44 | - | - | - |
| DP | Rule Design | 19.19 | 38.72 | 27.32 | - |
| **ODAO** | | **76.14** | **80.73** | **80.72** | **79.24** |

**Table 5: Results of *OTE* task from *OTE* module on SemEval dataset. We report the span-level $F_1$ scores on the test sets. Results of the baselines are reported from their original papers. - refers to unpublished results as of the date of writing (Feb. 2022).**

| Method | Human Effort | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|---|
| RINANTE | Gold Annotation | 81.96 | 85.67 | 72.09 | - |
| DeepWMaxSat | | 80.34 | 85.73 | - | 79.67 |
| DeepLogic | | 79.32 | 84.37 | - | 78.89 |
| **FS-ODAO** | | **85.47** | **87.23** | **84.56** | **88.43** |
| GMTCMLA | Sample Annotation | 67.10 | 78.70 | 64.37 | - |
| DP | Rule Design | 55.29 | 65.94 | 46.31 | - |
| **ODAO** | | **77.82** | **79.57** | **82.56** | **81.26** |

**Table 6: Results of *AOPE* task from *ASOE* and *OSAE* module on SemEval dataset. We report the span-level $F_1$ scores on the test sets. Results of the baselines are reported from their original papers. - refers to unpublished results as of the date of writing (Feb. 2022).**

| Method | Human Effort | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|---|
| QDSL | Gold Annotation | 70.20 | 78.05 | 71.22 | 77.28 |
| SDRN | | 67.13 | 76.48 | 70.94 | - |
| SpanMlt | | 68.66 | 75.60 | 64.48 | 71.78 |
| **FS-ODAO** | | **90.04** | **89.89** | **87.18** | **90.06** |
| **ODAO** | Rule Design | 81.75 | 83.02 | 83.93 | 81.41 |

raw corpus is provided as the input of ODAO. The self-training process is stopped if the count of pseudo-labeled reviews in a given iteration is less than 10.

## 4.5 Results and Discussion

To better compare the performance of different methods, we categorize them based on how much human effort is required.

**Results for ATE task:** The experimental results are shown in Table 4. The results show that ODAO significantly outperforms existing methods not trained on gold annotations and achieve competitive results compared to the fully supervised models. Moreover, the fully supervised version of the proposed method FS-ODAO also outperforms the state-of-the-art baselines for three out of the four datasets.

**Results for OTE task:** The experimental results are shown in Table 5. We can observe that DP performs much better on *OTE* tasks than on *ATE* tasks as the rules designed from DP are based on opinion lexicon. GMTCMLA also achieves better scores on *OTE* tasks but still has a significant gap compared with the fully supervised methods. ODAO significantly outperforms existing weakly supervised methods and achieves competitive results compared to fully supervised models. The fully supervised version FS-ODAO outperforms the state-of-the-art baselines on all four datasets with a big margin.

**Results for AOPE task:** The experimental results are shown in Table 6. *AOPE* is a more complex task than the aspect/opinion extraction tasks, as it requires the model to learn the correspondent relationships among the extracted terms. FS-ODAO not only outperforms all the baselines for all the datasets, but it also achieves better scores than those of the *ATE* and *OTE* tasks. It clearly shows the effectiveness of the double-layer design. The proposed ODAO even outperforms the state-of-the-art fully supervised baselines on two datasets ($\mathbb{S}_{14}l$ and $\mathbb{S}_{15}r$), illustrating that the model is well tolerated with noisy and biased weak supervision.

**Table 7: Ablation Study results showing $F_1$ score for span-level *ATE* task from *ATE* module on SemEval dataset.**

| Methods | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|
| **ODAO** | 76.14 | 80.73 | 80.72 | 79.24 |
| -Pair Extraction Modules | 50.13 | 57.53 | 60.86 | 60.71 |
| -Self Training | 62.06 | 72.19 | 72.13 | 71.0 |

**Table 8: Ablation Study results showing $F_1$ score for span-level *OTE* task from *OTE* module on SemEval dataset.**

| Methods | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|
| **ODAO** | 77.82 | 79.57 | 82.56 | 81.26 |
| -Pair Extraction Modules | 72.75 | 75.63 | 78.45 | 77.56 |
| -Self Training | 73.30 | 76.70 | 77.37 | 76.3 |

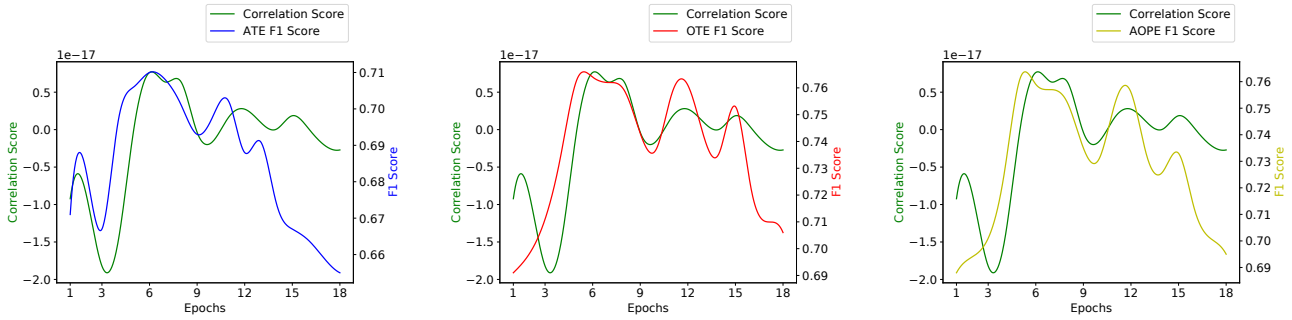
**Table 9: Ablation Study results showing $F_1$ score for span-level *AOPE* task combining *ASOE* and *OSAE* module on SemEval dataset.**

| Methods | $\mathbb{S}_{14}l$ | $\mathbb{S}_{14}r$ | $\mathbb{S}_{15}r$ | $\mathbb{S}_{16}r$ |
|---|---|---|---|---|
| **ODAO** | 81.75 | 83.02 | 83.93 | 81.41 |
| -Self Training | 70.64 | 76.21 | 76.65 | 76.09 |

## 4.6 Ablation Studies

We conduct ablation studies to investigate the contributions of each component to the overall model performances.

**Double-layer Design.** To illustrate the effectiveness of the double-layer design, we experiment with a single-layer architecture




(a) CCA score and ATE F1 score (b) CCA score and OTE F1 score (c) CCA score and AOPE F1 score

**Figure 3: CCA score and model performance on different tasks**

by removing *ASOE* and *OSAE* modules. Then the model conducts *ATE* and *OTE* tasks only. Since there are no modules to calculate the CCA score, we stop the training after a fixed number of epochs (= 5). For the self-training, to decide pseudo-label confidence, we consider a threshold of 1 for the start span ($h_{i_s} \geq 1$) and 1 for the end span ($h_{i_e} \geq 1$). We denote this as "-Pair Extraction Modules" and evaluate the model performance for *ATE* and *OTE* tasks.

The experimental results are shown in Table 7 and Table 8. We can observe that the model's performance drops significantly. The reasons are multi-fold: 1) Without the *ASOE* and *OSAE* modules, *ATE* and *OTE* modules are essentially trained independently. They cannot help each other with the tasks. 2) The idea of using a threshold to decide pseudo-label confidence does not provide a sufficient pseudo-labeled dataset to enrich the training set. 3) Furthermore, the pseudo-labeled dataset may contain higher noise due to model over-fitting, which can propagate with iterations.

**Self-training:** We adopt self-training to enrich the training data and reduce bias introduced due to the weak label generator. To validate the effectiveness of the self-training process, we perform experiments with the weak labels generated using the rules for model training only. We denote this setting as "-Self Training".

The results are shown in Table 7, Table 8, and Table 9 for the three tasks, respectively. We observe that the model performance drops for all three tasks significantly. These results indicate that the added weak labels are of high quality and the self-training process indeed enriches the training data.

**CCA as early stopping criteria:** We propose using *CCA* to evaluate the correlation between the related tasks and use it as an early stopping criterion. We claim that the correlation score can be used as an indicator for model fitness. To validate it, we train the model with the weak labels generated from the rules to show correlation score and model's performance for each training epoch.

Figure 3 shows the corresponding plots for *ATE*, *OTE*, and *AOPE* tasks. It can be observed from the plots that the correlation score and model performance are strongly related, and the model achieves high performance around the epochs where the correlation score is maximized. Specifically, the correlation score is maximized at epoch 6, and the model achieves higher performance for *ATE*, *OTE*, and *AOPE* tasks at epochs 6, 5, and 5, respectively. We can also observe that as epochs increase, the model's performance decreases due to over-fitting. This is also evident in the correlation score, which decreases as epochs increase.

## 4.7 Case Study

Table 10 shows prediction results by ODAO for some examples with complex aspect-opinion relation. There can be multiple pairs of aspect-opinion expressed in the same review, only aspect terms but no corresponding opinion term, one aspect term with multiple opinion terms, or multiple aspect terms with one opinion term. The proposed ODAO can handle all cases. Another interesting observation is that although the rules in the weak label generator restrict the aspect terms to be nouns and the opinion terms to be adjectives, ODAO can extract the verb "use" in the third example as aspect terms and the verb "recommend" in the first and last examples as opinions correctly.

# 5 CONCLUSION

This work proposes a double-layer span extraction framework to perform *ATE*, *OTE*, and *AOPE* tasks together for review analysis. To reduce the human effort for open-domain tasks, we proposed rules based on universal dependency parsing to label training data. The weak supervision is then used to train ODAO, a double-layer span extraction framework for aspect term extraction (ATE), opinion term extraction (OTE), and aspect-opinion pair extraction (AOPE) tasks. Canonical correlation analysis is used as an early stopping indicator to tackle the noise in the weak supervision so that the model will not over-fit to the noise. To tackle the bias issue of weak supervision, we propose enriching the training data by adding weak labels and conducting a self-training process. Extensive experiments demonstrate the power of the proposed ODAO. The results on four benchmark datasets for aspect-opinion co-extraction and pair extraction tasks show that ODAO can achieve competitive or even better performance compared with the state-of-the-art fully supervised methods. FS-ODAO, the fully supervised version of ODAO, achieves state-of-the-art performance and illustrates the double-layer design's effectiveness. Ablation studies show that ODAO can handle the noise and bias of the weak supervision.

**Table 10: Case study of reviews with complex aspect-opinion relation**

| Review | Model Predictions | Ground Truth |
|---|---|---|
| i recommend the black roasted codfish , it was the best dish of the evening . | **ATE**: [black roasted codfish, dish], **OTE**: [recommend, best], **AOPE**: [(black roasted codfish, recommend), (dish, best)] | **ATE**: [black roasted codfish, dish], **OTE**: [recommend, best], **AOPE**: [(black roasted codfish, recommend), (dish, best)] |
| – i ca n't say enough about this place . | **ATE**: [place], **OTE**: [null], **AOPE**: [(null, null)] | **ATE**: [place], **OTE**: [null], **AOPE**: [(null, null)] |
| it 's fast , light , and simple to use . | **ATE**: [use], **OTE**: [fast, light, simple], **AOPE**: [(use, fast), (use, light), (use, simple)] | **ATE**: [use], **OTE**: [fast, light, simple], **AOPE**: [(use, fast), (use, light), (use, simple)] |
| i can highly recommend their various saag and paneer and korma . | **ATE**: [saag, paneer, korma], **OTE**: [recommend], **AOPE**: [(saag, recommend), (paneer, recommend), (korma, recommend)] | **ATE**: [saag, paneer, korma], **OTE**: [recommend], **AOPE**: [(saag, recommend), (paneer, recommend), (korma, recommend)] |

## 6 ACKNOWLEDGEMENT

The work is supported in part by NIFA grant no. 2022-67015-36217 from the USDA National Institute of Food and Agriculture, and NSF IIS-2007941 from the National Science Foundation.

## REFERENCES

[1] Theodore Wilbur Anderson. 1962. *An introduction to multivariate statistical analysis.* Technical Report. Wiley New York.
[2] Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction. In *Proc. of ACL*. 6515–6524.
[3] Mark Craven and Johan Kumlien. 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proc. of ISMB*. 77–86.
[4] Hongliang Dai and Yangqiu Song. 2019. Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision. In *Proc. of ACL*. 5268–5277.
[5] Tijl De Bie and Bart De Moor. 2003. On the regularization of canonical correlation analysis. *Int. Sympos. ICA and BSS* (2003), 785–790.
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*. 4171–4186.
[7] Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proc. of NAACL*. 2509–2518.
[8] Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-Driven Span Labeling Model for Aspect–Opinion Pair Extraction. In *"Proc. of AAAI"*, Vol. 35. 12875–12883.
[9] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proc. of ACL*. 388–397.
[10] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*. Springer, 162–190.
[11] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of ACM SIGKDD*. 168–177.
[12] Arzoo Katiyar and Claire Cardie. 2016. Investigating lstms for joint extraction of opinion entities and relations. In *Proc. of ACL*. 919–929.
[13] JT Kent, John Bibby, and KV Mardia. 1979. *Multivariate analysis*. Academic Press Amsterdam.
[14] Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proc. of EMNLP*. 2886–2892.
[15] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proc. of SIGKDD*. 1054–1064.
[16] Ming Liao, Jing Li, Haisong Zhang, Lingzhi Wang, Xixin Wu, and Kam-Fai Wong. 2019. Coupling global and local context for unsupervised aspect extraction. In *Proc. of EMNLP-IJCNLP*. 4579–4589.
[17] Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. (2018).
[18] Abhishek Kumar Mishra and Mohna Chakraborty. 2021. Does local pruning offer task-specific models to learn effectively?. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*. 118–125.
[19] Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proc. of ACL*. 339–348.
[20] 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*. 19–30.
[21] 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)*. 486–495.
[22] 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*. 27–35.
[23] Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc. of HLT-EMNLP*. 339–346.
[24] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics* 37, 1 (2011), 9–27.
[25] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837.
[26] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning Named Entity Tagger using Domain-Specific Dictionary. In *Proc. of EMNLP*.
[27] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proc. of EMNLP-IJCNLP*. 5679–5688.
[28] Bo Wang and Houfeng Wang. 2008. Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing. In *Proc. of IJCNLP*.
[29] Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive Self-Training with Discriminator for Aspect Term Extraction. In *Proc. of EMNLP*. 257–268.
[30] Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proc. of ACL*. 2171–2181.
[31] Wenya Wang and Sinno Jialin Pan. 2020. Integrating deep learning with logic fusion for information extraction. In *Proc. of AAAI*, Vol. 34. 9225–9232.
[32] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In *Proc. of EMNLP*. 616–626.
[33] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proc. of AAAI*, Vol. 31.
[34] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *"Proc. of ACL"*. 1476–1488.
[35] Meixi Wu, Wenya Wang, and Sinno Jialin Pan. 2020. Deep Weighted MaxSAT for Aspect-based Opinion Extraction. In *Proc. of EMNLP*. 5618–5628.
[36] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In *Proc. of ACL*. 592–598.
[37] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *Proc. of NAACL-HLT*.
[38] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proc. of IJCAI*. 2979–2985.
[39] Jianfei Yu, Jing Jiang, and Rui Xia. 2018. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 1 (2018), 168–177.
[40] 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 *Proc. of ACL*. 3239–3248.
[41] Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proc. of NAACL*. 50–61.
[42] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proc. of CIKM*. 43–50.

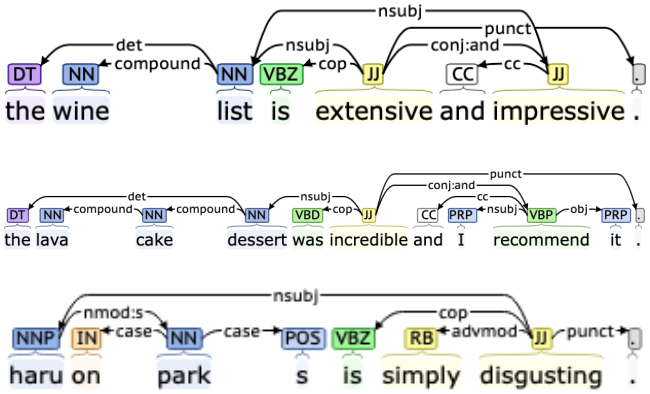



**Figure 4: Dependency Parse Tree Examples**

## A WEAK LABEL GENERATION

To generate weak labels for a review corpus, we design rules based on a dependency parsing tree. We first employ a universal dependency parser[4] to parse the review and obtain a dependency parse tree. Along with the dependency tree, we also obtain the part-of-speech tag information for the tokens in the review. As described in Section 3.3, we enrich the base rule with additional considerations. As a result, the following rules label aspect terms and opinion terms in the review.

(1) $AT = NN^* \leftarrow nsubj \leftarrow JJ^*(root) = OP$
(2) $OP = JJ^* \leftarrow comp \leftarrow OP$
(3) $AT = NN^* \leftarrow conj \leftarrow AT$
(4) $OP = JJ^* \leftarrow conj \leftarrow OP$
(5) $AT = NN^* \leftarrow comp \leftarrow AT$

All the used notations are explained in Table 11.

To properly extract the spans of aspect terms and opinion terms, we also utilize the phrase mining method [25]. This method extracts quality phrases using a quality phrase dictionary. We adopt the default dictionary provided by the tool[5], which is crawled from Wikipedia. The input to the phrase mining tool is the review corpus from the restaurant and laptop domains. The output is a ranked list of phrases with decreasing confidence scores. We obtain the top-ranked phrases with a confidence score greater than 0.9. If a review contains a top-ranked phrase and the rules already label some part of the phrase, then we consider the entire phrase. We only adjust the span boundary if all the tokens in the phrase are given the same label (either $AT$ or $OP$) using the rules.

| Notation | Definition |
|---|---|
| $OP$ | opinion term in the review |
| $AT$ | aspect term in the review |
| $NN^*$ | noun ∈ {singular or mass (NN), plural (NNS), singular proper noun (NNP), plural proper noun (NNPS)} |
| $JJ^*$ | adjective ∈ { adjective (JJ), comparative adjective (JJR), superlative adjective (JJS)} |
| $nsubj$ | nominal subject relation |
| $conj$ | conjunct relation |
| $comp$ | compound relation |

**Table 11: Summary of Notations**

Figure 4 presents three example reviews. For the first sentence, it can be observed that the word adjective *"extensive"* is connected to noun *"list"* via $nsubj$ relation. So, employing rule 1, *"extensive"* is labeled as an opinion term $OP$ and *"list"* is labeled as an aspect term $AT$. We can also observe that adjective *"impressive"* is connected to the opinion term *"extensive"* via $conj$ relation. So employing rule 4, *"impressive"* is labeled as another opinion term $OP$. Similarly, noun token *"wine"* is connected to the aspect term *"list"* via $conj$ relation. So employing rule 3, *"wine"* is labeled as aspect term $AT$.

For the second sentence, the adjective *"incredible"* is connected to noun *"dessert"* via $nsubj$ relation so by utilizing rule 1, *"dessert"* is labeled as an aspect term $AT$ and *"incredible"* is labeled as an opinion term $OP$. Also *"lava"* and *"cake"* are connected to *"dessert"* via $comp$ relation, therefore using rule 5, *"lava"* and *"cake"* are labeled as aspect terms $AT$.

For the third example, the adjective *"disgusting"* is connected to noun *"haru"* via $nsubj$ relation so by utilizing rule 1, *"haru"* is labeled as an aspect term $AT$ and *"disgusting"* is labeled as an opinion term $OP$. Applying the phrase mining results, *"haru on park s"* is a phrase, so the aspect term boundary is adjusted, and the entire phrase *"haru on park s"* is labeled as an aspect term $AT$ for the review.

[4]https://stanfordnlp.github.io/CoreNLP/depparse.html
[5]https://github.com/shangjingbo1226/AutoNER