Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis

Minh Hieu Phan & Philip Ogunbona
School of Computer Science and Software Engineering
University of Wollongong, NSW 2522, Australia
vmhp806@uowmail.edu.au, philipo@uow.edu.au

Abstract

The aspect-based sentiment analysis (ABSA) consists of two conceptual tasks, namely an aspect extraction and an aspect sentiment classification. Rather than considering the tasks separately, we build an end-to-end ABSA solution. Previous works in ABSA tasks did not fully leverage the importance of syntactical information. Hence, the aspect extraction model often failed to detect the boundaries of multi-word aspect terms. On the other hand, the aspect sentiment classifier was unable to account for the syntactical correlation between aspect terms and the context words. This paper explores the grammatical aspect of the sentence and employs the self-attention mechanism for syntactical learning. We combine part-of-speech embeddings, dependency-based embeddings and contextualized embeddings (e.g. BERT, RoBERTa) to enhance the performance of the aspect extractor. We also propose the syntactic relative distance to de-emphasize the adverse effects of unrelated words, having weak syntactic connection with the aspect terms. This increases the accuracy of the aspect sentiment classifier. Our solutions outperform the state-of-the-art models on SemEval-2014 dataset in both two subtasks.

1 Introduction

The process of understanding the sentiments expressed by consumers in a product review (opinionated text) is referred to as *sentiment analysis*. Deep insights into the opinionated text are gained through a fine-grained entity- or aspect-based sentiment labeling of the product being reviewed. Such insights can be invaluable for business decision making.

Aspect-based sentiment analysis (ABSA) consists of two sub-tasks, namely an aspect extraction (AE) and an aspect sentiment classification (ASC). However, the majority of reported works focused on one of the two sub-tasks alone. Representative works include (Xu et al., 2018; Da'u and Salim, 2019; Poria et al., 2016) for aspect extraction and (Zeng et al., 2019; Huang et al., 2018; Song et al., 2019; Thet et al., 2010) for aspect sentiment classification. Recent approaches (He et al., 2019; Wang et al., 2018; Li et al., 2019) attempted to develop an integrated solution to solve both tasks simultaneously by formulating both sub-tasks as a single sequence labelling with a unified tagging scheme. Adding unified tokens introduces overhead and complexity in the original ABSA tasks. Thus, multi-task models often have poorer performance compared with single-task models which are trained independently.

Recent advances in the NLU introduced contextualized language models, namely OpenAI GPT (Radford et al., 2018), BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). These models can capture the characteristics of word uses and account for different textual context in which words appear. Upon investigating the latest BERT/RoBERTa-based architectures used in aspect extraction, it became apparent that they were unable to determine the boundaries of multi-word aspects. For instance, the extractors broke the multi-word expression,"quality of food" into "quality of" and "food". We hypothesize that this shortcoming is caused by the inability of the contextualized embeddings to encode rich syntactical information.

In this paper, we integrate syntactical information into contextualized embeddings and propose an ABSA solution consisting of an aspect extractor and an aspect sentiment classifier as illustrated by Fig. 1. The proposed AE architecture, named *contextualized syntax-based aspect extraction* (CSAE), consists of POS embeddings, dependency-based embeddings (Levy and Goldberg, 2014) and self-attention in addition to RoBERTa layer.

Our ASC solution is closely related to the work

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3211–3220
July 5 - 10, 2020.

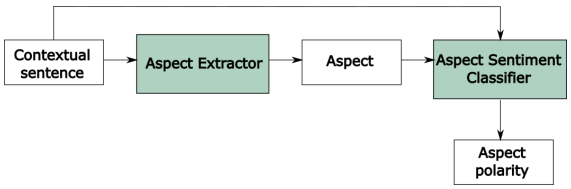


Figure 1: ABSA architecture

of Zeng et al. (2019) in which the *local context focus* (LCF) mechanism is exploited to down-weight the contribution of words that are far away from local context. However, this approach simply regarded the word counts between two words as their semantic relative distance and neglected the mutual syntactical relationship. Our approach employs the shortest path between two words in dependency parsing tree as a syntactic relative distance (SRD). We name this model *local context focus on syntax - ASC* (LCFS-ASC). Comparative experiments are conducted on two SemEval-2014 datasets (Pontiki et al., 2014) to demonstrate the importance of syntactical features in improving both AE and ASC models.

The main contributions of this paper can be highlighted as: (1) We propose the multi-channel CSAE model which distils grammatical aspects into contextualized features for improving sequential taggings; (2) We contribute the LCFS-ASC which can analyze syntactical connections between words to better understand local contexts that are relevant to target aspect terms; (3) We study the importance of the SRD by exploring the attention score in the LCF layer.

2 Related Work

This section details the evolution of ABSA solutions from word-embedding-based models to contextualized-embedding-based models and highlights their strengths and weaknesses.

Word-embedding-based Model

Recent ABSA works used pre-trained word embeddings as a data processing layer and added subsequent layers for a richer feature learning. Target-dependent Long Short-Term Memory (TD-LSTM) model (Tang et al., 2015) embedded the context words and target words into a vector space and employed LSTM cells to encode long-distance relationships in an input sequence. TD-LSTM captured the relatedness of target words with context words to extract relevant information for ABSA. Attention mechanism has been widely applied to the ABSA problem to overcome the vanishing gradients observed in long input sequence. Attention-based LSTM with Aspect Embedding (ATAE-LSTM) (Wang et al., 2016) utilized attention mechanism in addition to LSTM layers. Hence, the network can concentrate on crucial sentiment parts of a sentence in response to given aspects.

Contextualized Pre-trained Language Model

The quality of word representation is gauged by its capability to encode syntactical features and polysemic behaviour (i.e. word senses). Traditional word embeddings only produced single-context word representations. Recent works diverged from global word representations and considered context-dependent word embeddings which "described" the words differently in order to account for inherent word senses. BERT (Devlin et al., 2018) is a masked language model (LM) which masked a percentage of words in sentences and set up the training objective to predict the masked words. RoBERTa (Liu et al., 2019) improved upon BERT model by training the model longer with larger amount of data and eliminating next-sentence prediction objective. There have been several applications of BERT to the ABSA problem.

AEN-BERT (Song et al., 2019) used BERT to embed a context sequence and a target sequence; and applied attention to draw semantic interaction between targets and context words. LCF-BERT (Zeng et al., 2019) employed context dynamic masking/ context dynamic weighting to localize sentiment signals using semantic relative distance. This distance is measured by the word counts between the context word and target aspect terms. The local context layer allowed the model to emphasize semantic-relative contextual words. However, critical sentiment words sometimes can be associated with the target aspect terms through grammatical rules despite their large semantic relative distance. We hypothesize that using syntactical-relative-distance to identify unrelated words avoids mistakenly eliminating the contribution of crucial sentiment words.

There are examples of recent BERT-based approaches works that achieved promising results in AE tasks (see for example Xu et al. (2019)). However, they required re-training a BERT model on a large domain-specific corpus which made it infeasible to achieve a domain-independent aspect extractor. We abstain from such post-training ap-

proaches and look for a generic AE architecture.

3 Proposed Method

Given a contextual sentence S consisting of n tokens, $S = \{w_i|i \in [1,n]\}$, an end-to-end ABSA tasks aims to extract the set A of m aspect terms being mentioned where $A = \{a_i|i \in [1,m]\}$; and determine the polarity $y_p \in \{Positive, Negative, Neutral\}$ associated with each extracted aspect.

3.1 Aspect Extraction

Aspect extraction can be cast as a sequential labelling problem in which each input token w_i is assigned a label y_i. The labels y_i take on values from the set $\{B, I, O\}$ $(Begin, Inside, Outside)$, representing respectively the beginning of aspect term, inside of aspect term and the non-aspect tokens.

Fig. 2 depicts the overall architecture of the proposed contextualized syntax-based aspect extraction (CSAE) model. The CSAE consists of a contextualized embedding (e.g., BERT or RoBERTa), a part-of-speech embedding and a dependency-based embedding. The syntactical information in the final representation is enriched by concatenating the contextualized hidden states, attended POS states and attended dependency-based states.

3.1.1 Input Representation

The contextualized model requires a special classification token $[CLS]$ at the beginning of the input sequence and the separator $[SEP]$ appended to the end of input sequence. The input sentence is converted to the format "$[CLS]$" + Input sequence + "$[SEP]$".

3.1.2 Part-of-Speech Embedding

The part-of-speech (POS) of each word is annotated by the Universal POS tags [1]; subsequently the POS of an input sequence $P = \{p_1, p_2, ..., p_n\}$ is retrieved. The POS embedding layer takes the sparse vector representation P to extract a dense vector representation $V^P = \{v_i^p|i \in [1,n]\}$ wherein $v_i^p \in R^{h\_pos\_emb}$, and h_pos_emb refers to the hidden size of the POS embeddings. Then, the self-attention layer is utilized to observe the entire sequence of POS taggers and extract the grammatical dependencies in the input sentence.

[1] Universal POS Tags. URL: https://universaldependencies.org/u/pos/

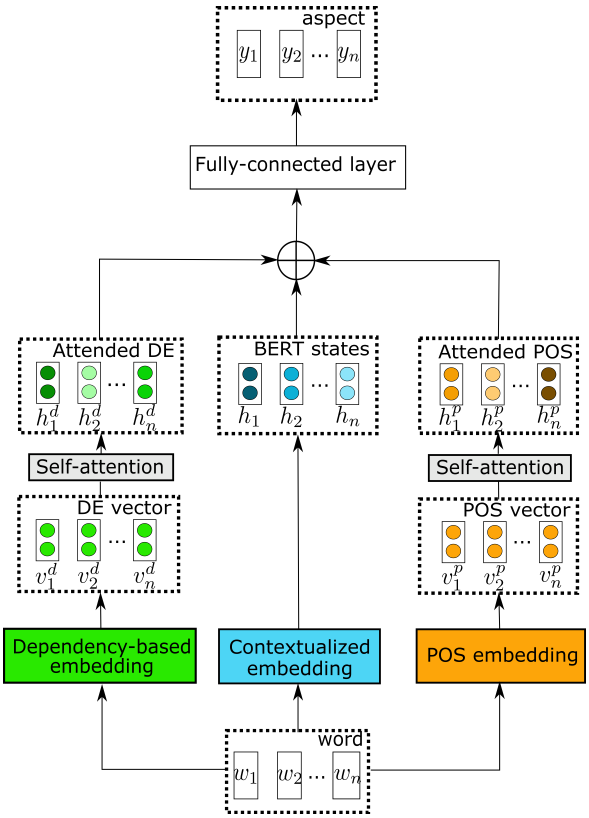
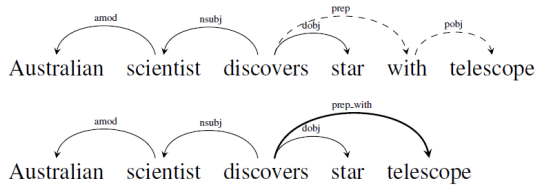


Figure 2: Overall architecture of the proposed CSAE

3.1.3 Dependency-based Embedding

Instead of using a linear bag-of-words context to form a context window, the dependency-based embedding (Levy and Goldberg, 2014) (DE) uses dependency-based contexts based on the syntactical relations in which the word participates. The process starts by using a dependency tree to parse the sentence. For each target word w and the modifiers $m_1, m_2, \ldots, m_n$ associated with w, the context $C = \{(m_1, rel_1), (m_2, rel_2), \ldots, (m_n, rel_n)\}$ is constructed. In this consideration, rel_i is the dependency relation (e.g., *subj, amod, pobj*) between a target word w and a modifier m_i, while rel^{-1} represents the inverse relations. Before extracting the final contexts, the relations consisting of a preposition are collapsed by subsuming the preposition into a dependency label. Fig. 3 describes the process of collapsing prepositions into a dependency relation and demonstrates the extracted contexts of each target word in a given sentence. The DE can incorporate the distant relation which is out of reach in linear-context word embedding. It also de-emphasizes irrelevant words accidentally falling into the context windows.



WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Figure 3: Dependency-based context example. **Top**: prepositions are collapsed into a single arc, *telescope* is a direct modifier of *telescope*. **Bottom**: contexts extracted for each word in a sentence

3.1.4 Fine-tuning Procedure

The training objective is to minimize the cross-entropy loss with $\mathcal{L}_2$ regularization. Specifically, the optimal parameters θ of the deep learning model are obtained from

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \hat{y}_i \log y_i + \lambda \sum_{\theta \in \Theta} \theta^2, \qquad (1)$$

where λ is the regularization parameter and $\hat{y}_i$ the predicted label corresponding to y_i.

3.2 Aspect Sentiment Classification

Given a contextual sentence $S = \{w_i | i \in [1, n]\}$ and extracted aspect terms $A = \{a_i | i \in [1, m]\}$, we need to determine the polarity $\{Positive, Neutral, Negative\}$ of the aspect terms in the contextual sentence.

Fig. 4 illustrates the overall architecture of the proposed Local Context Feature-Aspect Sentiment Classification including two independent Contextualized Embedding for global and local contexts.

3.2.1 Input Representation

To comprehend the global context, the contextual sentence S and aspect terms A are combined to construct global contexts G. The input format of global context G is $G = [CLS]+S+[SEP]+A+[SEP]$. On the other hand, the local contexts L is the contextual sentence S whose format is $[CLS] + S + [SEP]$. In BERT architecture, the global context G is explicitly represented as a pair of text consisting of a contextual sentence S and aspect terms A. When a token in G belongs to a first or second segment of the sentence pair, its segment token is

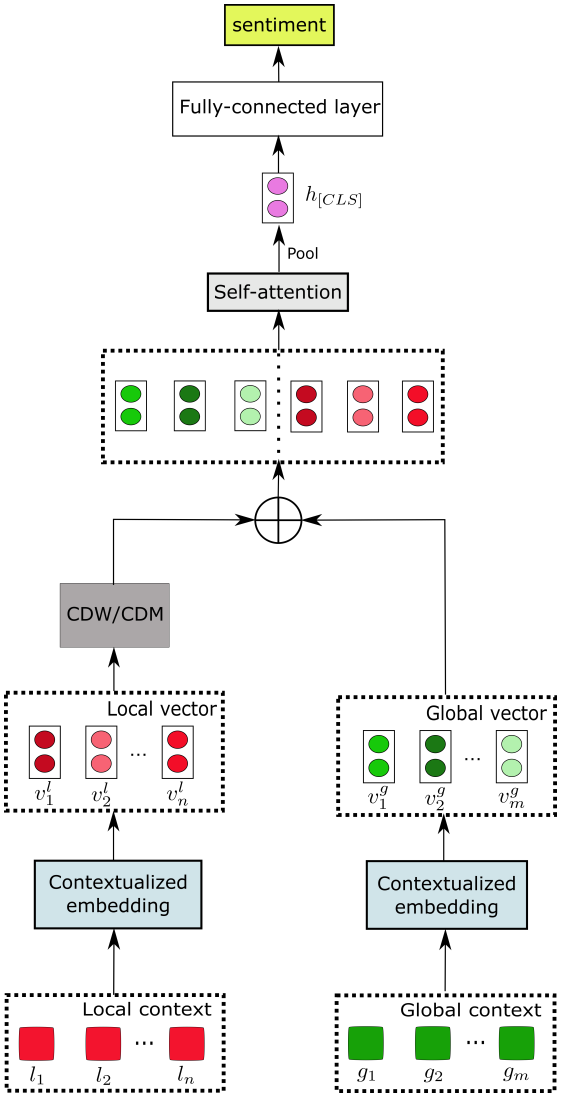


Figure 4: Overall architecture of the proposed LCF-ASC

indexed as 1 or 2 respectively. This next-sentence-prediction characteristic of the BERT model allows BERT-based ASC models to capture the semantic relationship between the contextual sentence and the aspect. Since RoBERTa removed the next-sentence-prediction task when training the model, it is suspected that the RoBERTa representation is not as informative as the BERT representation for the ASC task. The hidden state corresponding to a special classification token $[CLS]$ represents the aggregation of the entire sentence.

3.2.2 Local Context Focus

The local context vectors $V^l = \{v_i^l | i \in [1, n]\}$ are obtained by feeding the local contexts into the contextualized embedding. Next, we apply context feature dynamic weight/context feature dynamic mask (CDW/CDM) (Zeng et al., 2019) techniques

on V^l to alleviate the negative influence of irrelevant opinion words which are distant from the target aspect terms.

Relative Distance

The SRD between words is measured by the shortest distance between their corresponding nodes in the dependency-parsed tree. If the aspect term is composed of multiple words, the SRD between an input word and a multi-word aspect term is computed as an average distance between each component word and an input word. Fig. 5 illustrates the dependency-parsed tree constructed from a sample product review. The SRD between an aspect term "sound amplifier" and sentiment word "loudly" is computed as:

$$\begin{aligned} &\text{SRD}(\text{amplifier}, \text{loudly}) = 2 \\ &\text{SRD}(\text{sound}, \text{loudly}) = 3 \\ \implies &\text{SRD}(\text{sound amplifier}, \text{loudly}) = 2.5. \end{aligned}$$

On the other hand, the semantic relative distance when counting words between "sound amplifier" and "loudly" is 7 (as demonstrated in (Zeng et al., 2019)) which might make key sentiment words being down-weighted undesirably.

Context dynamic mask (CDM) masks out the less-semantic context features whose SRD to target words is greater than the pre-defined threshold. Given the local contexts V^l, the mask vector V_i^m for each contextual word m_i is computed based on certain SRD threshold α:

$$\begin{aligned} v_i^m &= \begin{cases} O & SRD_i > \alpha \\ I & SRD_i \leq \alpha \end{cases} \\ M &= [v_1^m, v_2^m, ..., v_n^m] \\ V^{CDM} &= V^l \odot M \end{aligned} \tag{2}$$

O and I are vectors of all zero and one respectively; O and $I \in R^h$ where h is the hidden size of a contextualized embedding and also the dimension of local context vector v_i^l. $\odot$ represents the element-wise dot product to mask out the local vector V^l by using the mask matrix M

Context dynamic weighting retains the contribution of less-semantic-relative context features but de-emphasizes them based on their distance to aspect terms. Thus,

$$\begin{aligned} v_i^w &= \begin{cases} (1 - \frac{SRD_i - \alpha}{N}) \cdot I & SRD_i > \alpha \\ I & SRD_i \leq \alpha \end{cases} \\ W &= [v_1^w, v_2^w, ..., v_n^w] \\ V^{CDW} &= V^l \odot W \end{aligned} \tag{3}$$

where N is the length of the contextual sentence.

Fine-tuning Procedure

The hidden state of classification token "[CLS]" h_{pool} is pooled out and fed into a softmax layer to predict the polarity from the set {*Positive, Neutral, Negative*}. Similarly to the AE model, we use the cross-entropy loss with $\mathcal{L}_2$ regularization as a loss function to fine-tune the entire ASC deep-learning model.

4 Performance Evaluation

4.1 Dataset

We evaluate and compare the proposed AE and ASC models on two benchmark datasets as described in Table 1. They are laptop-domain and restaurant-domain datasets taken from SemEval-2014 Task 4 challenge (Pontiki et al., 2014). Each sample sentence in the datasets is annotated with marked aspect terms and their associated polarity.

Table 1: Number of instances by polarity in training and test data

Dataset	Training			Testing		
	Pos	Neg	Neu	Pos	Neg	Neu
Restaurant	1315	462	368	426	143	146
Laptop	602	514	260	201	197	94

4.2 Baseline Models

We benchmark the performance against recent models in ABSA tasks to demonstrate the effectiveness of the proposed CSAE model and LCFS-ASC model.

The **first** group of models follow pipelining approach which train single-task models independently and pipeline the output of AE and ASC to build an end-to-end ABSA solution. To highlight the improved performance of the contextualized embeddings in ABSA tasks, we pick top high-performing word-embedding-based and contextualized-embedding-based models in both AE and ASC tasks. For a fair comparison, we only consider domain-independent models and eschew comparing with post-training approaches because they require re-purposing the entire model on large corpora before fine-tuning it for the in-domain end task.

For *AE* task, we select two word-embedding-based model and one contextualized-embedding-based model to demonstrate that a simple BERT

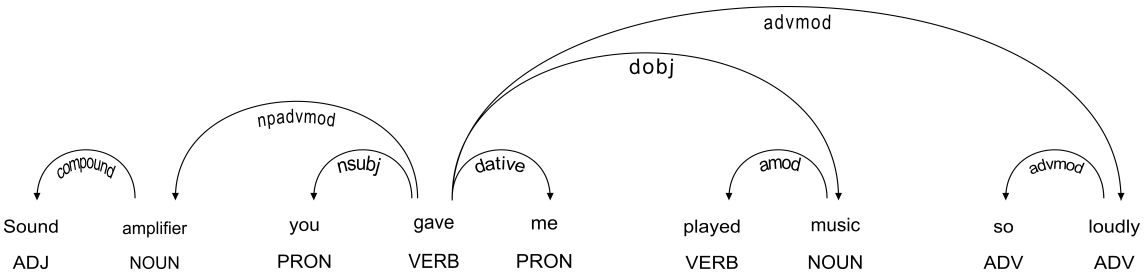


Figure 5: Dependency-parsed tree of the product review

layer can outperform a sophisticated network using word embeddings:

BiLSTM (Liu et al., 2015) is a Named Entity Recognition model employing Bidirectional LSTM on top of a Word Embedding representation.
DTBCSNN (Ye et al., 2017) is a dependency tree based stacked convolutional neural network which used the inference layer for aspect extraction.
BERT-AE (Devlin et al., 2018) utilizes a BERT representation for AE. This model acts as a reference to demonstrate the importance of our designed components adding to a contextualized representation.

For *ASC* task, we select two word-embedding-based models and four contextualized-embedding-based models. Various BERT-based models are examined to demonstrate that the provided information about aspects can be employed to attend to relevant sentiment information and improve the BERT-based ASC models:

AOA (Huang et al., 2018) uses multiple attention layers to model the interaction between aspects and sentences.
MGAN (Fan et al., 2018) uses fine-grained and coarse-grained attention to capture word-level interaction between aspects and sentences.
BERT-ASC (Devlin et al., 2018), utilizes a BERT representation for ASC
BERT-PT (Xu et al., 2018) re-trains a contextualized BERT model on a large domain-specific corpus to enhance the quality of word representations to the end-task.
AEN-BERT (Song et al., 2019) adopts contextualized BERT model and attention mechanism to model the relationship between context and targets. This model is used to show the improvements in ASC tasks when leveraging additional information about target terms in the given context.
LCF-BERT (Zeng et al., 2019) employs Local-Context-Focus design with Semantic-Relative-Distance (SeRD) to discard unrelated sentiment words. This model acts as a reference to illustrate the importance of our proposed SRD metrics in improving ASC models. Since the choice of BERT model is not indicated in the paper (Zeng et al., 2019) and we do not have an access to $BERT_{large}$ model, we re-implement the LCF-BERT model using the $BERT_{base}$ model based on their proposed methodology.

The **second** group consists of integrated approaches which aim to extract aspect terms and determine polarity simultaneously through a unified tagging scheme. This group of models can model the joint information in both sub-tasks and leverage all available sources of training information to handle an end-to-end ABSA problem:

MNN (Wang et al., 2018) employs attention mechanism to jointly learn the relationship between aspects and sentiments for a multi-task neural network.
UABSA (Li et al., 2019) is a unified model for ABSA, consisting of two stacked RNNs for the target boundary detection tasks (auxiliary) and the complete ABSA tasks (primary).
IMN (He et al., 2019) uses message passing architecture to transfer information iteratively through different tasks along latent variables.

4.3 Model Variations

To evaluate our proposed models along with their components in both AE and ASC tasks, we conduct a series of experiments with different settings.

For our proposed *AE* solution, we perform ablation study where certain modules are removed from the CSAE architecture to show their effects on the end performance:

RoBERTa-AE utilizes a RoBERTa representation to demonstrate the improved quality of the RoBERTa representation in AE task.
RoBERTa-POS employs a RoBERTa representation and a POS embedding to demonstrate that POS is helpful to identify aspect terms in a sentence.

RoBERTa-Dep uses a RoBERTa representation and a dependency-based embedding to compare the effects of dependency-based features and POS features in AE tasks.
CSAE is a complete model, consisting of RoBERTa, POS embedding and dependency-based embedding layers.

For our proposed *ASC* solution, we experiment with the RoBERTa-ASC model without the LCF layer and a complete LCFS-ASC model with the LCF layer. Hence, the impact of LCF layer on ASC tasks can be demonstrated.

RoBERTa-ASC utilizes a RoBERTa representation for ASC to compare the suitability of BERT and RoBERTa representations in ASC tasks.
LCFS-ASC-CDW is a LCFS-ASC model employing CDW technique.
LCFS-ASC-CDM is a LCFS-ASC model employing CDM technique.

Note that we used the $BERT_{base}$ to implement LCFS-ASC model due to the lack of adequate computing resources, as well as to ensure the fair comparison between the LCF-BERT and our proposed model. Similarly, the CSAE model is built on top of the $RoBERTa_{base}$ model. For AE task, we use the standard evaluation script provided by SemEval challenge to report F1-score. On the other hand, the accuracy and macro F1-score over 3 classes of polarities are considered to be evaluation metrics for ASC task.

5 Experiments

Table 2: The examples column shows the sentences having multi-word aspect terms being highlighted in red. The two following columns display the predicted aspect terms by RoBERTa-AE and CSAE models respectively

Examples	RoBERTa-AE	CSAE
1. Try the Times Square cocktail – ginger lemonade with vodka (also available without vodka)	cocktail	Times Square cocktail
2. The restaurant offers no desserts beyond the complimentary espresso cup filled with chocolate mousse	espresso cup filled with, chocolate mousse	espresso cup filled with chocolate mousse
3. Then just the other day, my left "mouse" button snapped!	"mouse" button	left "mouse" button

Table 2 compares the performance of the RoBERTa-AE-based model and the complete CSAE model. It is noticeable that the CSAE model outperforms RoBERTa-AE model in defining the boundary of multi-word aspect terms. Using a contextualized RoBERTa feature, the RoBERTa-AE is only able to identify the noun "cocktail" in a noun phrase, suggesting a RoBERTa representation fails to capture rich syntactical structure in a contextual sentence. In the universal dependencies schema, "Times" and "Square" are a *PROPN* (proper noun) tag which is part of the name of specific place, and have *compound* relation with the noun "cocktail". Being given explicit information about special syntactical properties of an example, CSAE successfully identifies a compound noun as an aspect term even though an aspect term "Time Square cocktail" does not appear in a training set. Additionally, even though RoBERTa-AE can identify individual aspect terms "espresso cup filled with" and "chocolate mousse" in example 2, it fails to group them together to form a complete multi-word term. CSAE, on the other hand, is able to model the role of the preposition "with" and detect the true boundary of the aspect term.

6 Results & Analysis

6.1 Aspect Extraction

6.1.1 Main Results

Table 3 summarizes the results of our proposed models compared with the baseline models. When compared with the word-embedding-based models, our CSAE model performs better than the BiLSTM and DTBCSNN models with gains of 3.93 percentage points (p.p), 1.99p.p and 5.23p.p, 2.68p.p in laptop and restaurant datasets respectively. The performance of our model is close to IMN's in laptop domain and outperforms other integrated approaches in both settings. Especially, our CSAE model has F1-score at least 3.32 p.p higher than other integrated approaches in the restaurant domain, suggesting that single-task models can significantly outperform integrated solutions with sophisticated architecture by simply improving the quality of feature representations.

6.1.2 Ablation Study

To investigate the effects of different designed components in a CSAE, we start with a base model using just a RoBERTa representation for aspect extraction and add other components one at a time. We found that our base model always gives superior performance compared to the BERT-based model. The performance is improved when we introduce

the POS embedding and dependency-based embedding to capture rich syntactical information. The POS embeddings solely represent the POS of each individual word and leave the feature extraction job for the attention layer, while the dependency-based embeddings directly infuse the grammatical interaction between words into the word representation. Hence, it is expected that RoBERTa with dependency-based features has slightly higher F1-score than RoBERTa with POS features. Overall, CSAE with full complement of both components gained significant improvement. It suggests that the RoBERTa model has not entirely "comprehended" the grammatical aspects of natural language and there is room for improvements in contextualized LM by further leveraging syntactical information of sentences.

Table 3: Comparison of our best performing AE model variants in terms of F1 scores (%) with the state-of-the-art methods

	Domain	Laptop	Rest
	Model	**F1**	**F1**
Single-task	**BiLSTM**	73.72	81.42
	DTBCSNN	75.66	83.97
	BERT-AE	73.92	82.56
Integrated	**MNN**	76.94	83.05
	UABSA	77.34	83.92
	IMN	**77.96**	83.33
Proposed	**RoBERTa-AE**	75.22	85.12
	RoBERTa-POS	76.01	85.56
	RoBERTa-Dep	76.88	86.25
	CSAE	77.65	**86.65**

Note: The best result in each dataset is highlighted in bold

6.2 Aspect Sentiment Classification

6.2.1 Main Results

Table 4 demonstrates that our proposed LCFS-ASC using Syntactic Relative Distance to localize the context features has the best performance in both Laptop and Restaurant dataset. The single-task, integrated and our proposed approach are displayed in the first, second and third parts, respectively. Our proposed model outperforms the BERT-PT by a large margin without utilizing additional knowledge from a larger corpus to train domain-specific embeddings. All BERT-based single-task models outperform the integrated models, suggesting that the unified tagging schema imposed overheads to the ASC tasks by introducing extra classes. As discussed in Section 3.2.1, the removal of the next-sentence-pair task in RoBERTa makes the RoBERTa representation less suitable to the ASC task leading to the underperformance of RoBERTa-ASC.

Table 4: Comparison results of our best performing ASC model variants in terms of F1 scores and accuracy (%) with the state-of-the-art methods

Domain	**Laptop**		**Rest**	
Model	**F1**	**Acc**	**F1**	**Acc**
AOA	-	74.5	-	81.2
MGAN	72.47	75.39	71.94	81.25
BERT-ASC *	72.68	76.25	76.98	84.46
BERT-PT	75.08	78.07	76.96	84.95
AEN-BERT	76.31	79.93	73.76	83.12
LCF-BERT-CDW *	76.20	80.21	79.12	85.91
LCF-BERT-CDM *	75.76	79.65	78.74	85.73
MNN	65.98	70.40	68.45	77.17
UABSA	68.24	72.30	68.38	79.68
IMN	72.02	75.36	75.66	83.89
RoBERTa-ASC	70.52	74.12	75.12	82.82
LCFS-ASC-CDW	**77.13**	**80.52**	**80.31**	**86.71**
LCFS-ASC-CDM	76.45	80.34	80.10	86.13

Note: The best result in each dataset is highlighted in bold. The results of models we reproduced by following the methodology published in the paper are indicated by asterisk (*).

The proposed LCFS-ASC has a slightly improved performance compared with the LCF-BERT when using either CDM or CDW. The result demonstrates the effectiveness of Syntactical Relative Distance in encoding syntactical information. CDW helps to boost the performance of LCFS-ASC model more than the CDM. Since CDM completely blocks the signals of the contexts being identified unimportant, it may falsely disregard useful signals. On the other hand, CDW emphasizes flexibility and allows further signals to contribute small weights corresponding to its relatedness with the aspect terms in the dependency-based tree.

6.2.2 Analysis of SRD's Effects by Visualizing Attention Scores

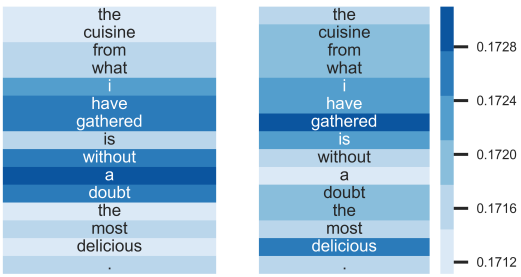


Figure 6: Attention scores of LCF-BERT-CDW (left) and LCFS-ASC-CDW (right)

Fig. 6 visualizes the attention score for the best-

performing LCFS-ASC-CDW and LCF-BERT-CDW models. For a given input sentence, LCFS-ASC assigns a correct *positive* polarity to the aspect term "cuisine", while LCF-BERT gives a wrong prediction as *negative*. Since LCF-BERT uses Semantic Relative Distance, the sentiment term "without a doubt" has been paid the most focus due to its close distance to the aspect term "cuisine" based on word counts metrics. On the other hand, the signal of a key sentiment word "delicious" is mistakenly down-weighted because it is far away from the aspect term "cuisine". Nevertheless, the LCFS-ASC retains the importance of the word "delicious" because Syntactical Relative Distance accounts for the direct interaction between the adjective "delicious" and the aspect term "cuisine" in a dependency-based tree.

7 Conclusion and Future work

We proposed an end-to-end ABSA solution which pipelined an aspect extractor and an aspect sentiment classifier. The results indicate that exploitation of syntactical structures of sentences empowers the contextualized models to improve on current works in both ASC and AE tasks. Our proposed aspect sentiment classifier outperformed post-training ASC model and enabled the creation of a domain-independent solution. The proposed SRD allows the aspect sentiment classifier to focus on critical sentiment words which modify the target aspect term through dependency-based structure. The substantial improvements highlight the under-performance of recent contextualized embedding models in "understanding" syntactical features and suggests future directions in developing more syntax-learning contextualized embeddings. One can try to adapt our proposed CSAE architecture for an integrated approach by applying the unified tagging scheme; thereby, aspect extraction and sentiment classification can be achieved simultaneously.

8 Acknowledgement

Thanks to Vinh Hung Ngo, who has provided insightful advice to improve my writings and experimental results.

References

Aminu Da'u and Naomie Salim. 2019. Aspect extraction on user textual reviews using multi-channel convolutional neural network. *PeerJ Computer Science*, 5:e191.

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

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1906.06906*.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. *arxiv preprint arXiv:1804.06536*.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

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

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.

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

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Hai Ye, Zichao Yan, Zhunchen Luo, and Wenhan Chao. 2017. Dependency-tree based convolutional neural networks for aspect term extraction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 350–362. Springer.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.