Syntactically Aware Cross-Domain Aspect and Opinion Terms Extraction

Oren Pereg, Daniel Korat, Moshe Wasserblat
Intel AI Lab, Petah Tikva, Israel
{oren.pereg, daniel.korat, moshe.wasserblat}@intel.com

Abstract

A fundamental task of fine-grained sentiment analysis is aspect and opinion terms extraction. Supervised-learning approaches have shown good results for this task; however, they fail to scale across domains where labeled data is lacking. Non pre-trained unsupervised domain adaptation methods that incorporate external linguistic knowledge have proven effective in transferring aspect and opinion knowledge from a labeled source domain to unlabeled target domains; however, pre-trained transformer-based models like BERT and RoBERTa already exhibit substantial syntactic knowledge. In this paper, we propose a method for incorporating external linguistic information into a self-attention mechanism coupled with the BERT model. This enables leveraging the intrinsic knowledge existing within BERT together with externally introduced syntactic information, to bridge the gap across domains. Finally, we demonstrate enhanced results on three benchmark datasets.

1 Introduction

A fundamental task of fine-grained sentiment analysis is aspect and opinion terms extraction. For example, in the sentence *"The chocolate cake was incredible"*, the aspect term is *chocolate cake* and the opinion term is *incredible*. Most of the work related to aspect and opinion term extraction is formulated as a supervised sequence-tagging task. RNN-based models (Liu et al., 2015) and Transformer-based models showed promising results in single-domain setups where the training and the testing data are from the same domain. However, these approaches typically do not scale across different domains, where only unlabeled data is available for the target domain, since aspect terms from two different domains are usually semantically different hence separated in the embedding space. For example, frequent aspect terms in the restaurant domain, like *salad* and *dessert*, have little or no semantic relatedness to frequent aspect terms in the laptop domain, like *screen size* and *battery life*. To date, only a handful of approaches for unsupervised domain adaptation of aspect and opinion term extraction have been proposed.

It has been shown that syntactic information is important for identifying aspect and opinion terms (Hu and Liu, 2004b; Qiu et al., 2011). A recent line of work, based on non pre-trained models, encodes dependency-based aspect extraction rules (Ding et al., 2017) or automatically-generated dependency relations (Wang and Jialin Pan, 2018; Wang and Pan, 2020), as auxiliary supervision for non pre-trained models. This recent line of work demonstrates effective domain adaptation by incorporating syntactic knowledge into non pre-trained models during their training step. Subsequently, recent studies (Clark et al., 2019; Htut et al., 2019) show that pre-trained transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) already exhibit substantial linguistic knowledge. In this paper we examine whether the incorporation of external syntactic knowledge into pre-trained models, contributes to bridging the gap across domains. For this purpose, we propose an approach for unsupervised domain-adaptation of aspect and opinion terms extraction based on incorporating linguistic knowledge into a pre-trained BERT model. Specifically, inspired by Strubell et al. (2018), we incorporate externally-generated dependency relations into a self-attention mechanism that is coupled with the pre-trained BERT model

Proceedings of the 28th International Conference on Computational Linguistics, pages 1772–1777
Barcelona, Spain (Online), December 8-13, 2020

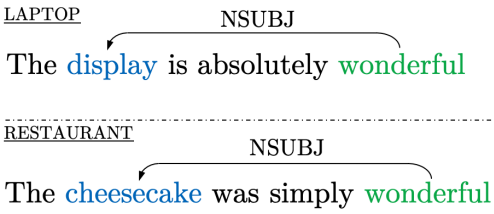

Figure 1: An example of opinionated sentences from two different domains with similar syntactic patterns. Opinion terms are colored green and aspect terms are colored blue.

(Stickland and Murray, 2019), where the external information is introduced during the fine-tuning and testing stages of the model.

2 Motivation and Background

Formally, the task of aspect and opinion terms extraction can be formulated as a sequence tagging task. The input is a sequence of tokens $X = \{x_1, x_2, .., x_n\}$ where the objective is to predict a corresponding sequence of labels $Y = \{y_1, y_2, ..., y_n\}$ with $y_i \in \{BA, IA, BO, IO, N\}$, where BA, BO, IA and IO represent a beginning of aspect/opinion and inside of aspect/opinion, respectively, and N represents all other tokens. The goal of unsupervised domain adaptation is to predict the token-level labels y_i^T of unlabeled target domain sentences $D^T = \{(x_i^T)\}$, given a set of labeled sentences from a source domain $D^S = \{(X_j^S, Y_j^S)\}$.

It was observed that aspect and opinion terms maintain often-occurring syntactic patterns (Hu and Liu, 2004b; Qiu et al., 2011). Consider for example, a sentence from the laptop domain *"The display is absolutely wonderful"* and a sentence from the restaurant domain *"The cheesecake was simply wonderful"*. In the first sentence, an NSUBJ dependency relation exists between the opinion term (*'wonderful'*) and the aspect term (*'display'*). Assuming that the pattern aspect-NSUBJ-opinion is frequently observed in the laptop domain, then the term *cheesecake* can be extracted as an aspect term in the restaurant domain (Figure 1). This domain-independent trait of the syntactic structure can be leveraged for transferring knowledge from a labeled source domain to an unlabeled target domain. Based on this notion, Ding et al. (2017) proposed using dependency-based aspect extraction rules as auxiliary supervision for an RNN network. However, this method depends on the quality of manually-crafted rules. Wang and Jialin Pan (2018) addressed this issue by automatically encoding dependency relations into the hidden representations of words, thus shifting the representations of different aspect terms having similar dependency relations, close to each other. Wang and Pan (2020) have further enhanced this model by integrating a conditional domain-adversarial network that encodes both word features and syntactic parent relation types.

Analyses of pre-trained transformer-based models like BERT reveal substantial syntactic information captured within their attention mechanisms; however, those analyses also show that for many syntactic relations BERT only slightly improve over a simple baseline (Clark et al., 2019; Htut et al., 2019). Our goal is to design a neural network model that leverages both the information captured in the pre-trained model, and externally introduced syntactic information, to bridge the gap between the source and target domains.

3 The Proposed Model

The basis for our model is a pre-trained BERT-base model (Devlin et al., 2019) with a fully connected sequence tagging classifier on top. Inspired by the work of Strubell et al. (2018), we incorporate dependency relations into a self-attention mechanism denoting a syntactically-aware attention head. Our approach differs from previous approaches which modify an existing self-attention head within a transformer-based model and train it from scratch. Our method modifies the BERT function by adding syntactically-aware self-attention heads in parallel to the BERT model (Stickland and Murray, 2019), and introduce the syntactic knowledge during the fine-tuning and testing stages. This leaves the original

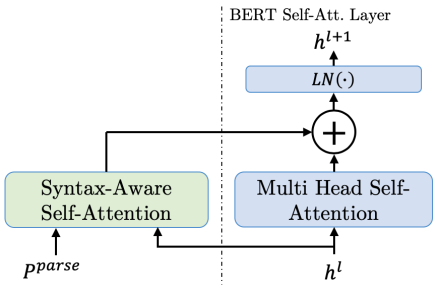

Figure 2: Coupling a syntactically-aware self-attention with a multi-head self-attention layer in a BERT model.

pre-trained model intact, enabling the model to utilize both the external linguistic information that is incorporated into the model and the intrinsic knowledge gained during the pre-training stage of the model. We refer to this model as syntactically-aware extended attention layer (SA-EXAL).

Multi-Head Self-Attention. The basis of our implementation is BERT's multi-head self-attention mechanism (Vaswani et al., 2017), which consists of I scaled dot-product attention heads. For each attention head i, the hidden token representations $h^l \in R^{d \times T}$, at the input of layer l, are projected to key, query and value representations K_i, Q_i and V_i of dimensions $T \times d_k$, where T is the number of tokens in the input sequence and $d_k = d/I$. Attention head i denotes attention weights that are a distinct distribution of every input token over all other tokens in the sequence:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}) \tag{1}$$

The output of attention head i is denoted by $M_i = A_i V_i$, where M_i is a $T \times T$ matrix, in which each row t, represents a weighted sum of the value representations of all other tokens with respect to token t. Finally, the outputs of all I attention heads are concatenated and projected through a feed-forward(FF) network: $SA = FF(M_1, M_2, ..., M_I)$.

Syntactically-Aware Self-Attention. Inspired by the work of Strubell et al. (2018), we incorporate syntactic information into the self-attention head, forming a syntactically-aware self-attention, by encouraging it to attend to specific tokens corresponding to the syntactic structure of the sentence. As in the original attention-heads, we project h^l denoting K_{parse}, Q_{parse} and V_{parse} matrix representations of dimensions $T \times d_k$, but unlike the original heads, we also use an external syntactic parser (Dozat and Manning, 2017) to generate P_{parse}, a $T \times T$ matrix in which each row t represents the probability of each token in the sentence to be the syntactic head of token t. We encourage this self-attention head to attend to the syntactic head of each token by performing an element-wise multiplication between P_{parse} and the dot product between the key and query matrices:

$$A_{parse} = softmax(\frac{(Q_{parse} K_{parse}^T) * P_{parse}}{\sqrt{d_k}}) \tag{2}$$

As in the original heads, The output of the syntactically-aware self-attention head is denoted by: $SA_{parse} = FF(A_{parse} V_{parse})$.

Adding Syntactically-Aware Self-Attention to BERT. Inspired by the work of Stickland and Murray (2019) we modify the BERT$(\cdot)$ function by adding a syntactically-aware self-attention head in parallel to each self-attention layer of the BERT model (see Figure 2) as follows:

$$h^{l+1} = LN(h^l + SA(h^l) + SA_{parse}(h^l)) \tag{3}$$

where LN is BERT's layer normalization function and $h^l \in R^{d \times T}$ are the T hidden token representations at the input of layer l. Note that the contribution of the $SA_{parse}(h^l)$ component to the representation of

each token t in layer $l+1$, is mostly the representation of the syntactic head of token t. This shifts the representations of aspect terms from distinct domains, that syntactically relate to the same opinion term, closer to each other, thus contributing to bridging the gap between the domains.

4 Experiments

Data & Experimental Setup. Our experimental setup follows that of Wang and Pan (2020). We conduct experiments on benchmark datasets with customer reviews from three different domains: restaurant, laptop and digital devices. The restaurant domain combines reviews from SemEval 2014 (Pontiki et al., 2014) and SemEval 2015 (Pontiki et al., 2015). The laptop domain contains laptop reviews from SemEval 2014. Opinion term labels for these domains are obtained from Wang et al. (2016). For the device domain, we use reviews from Hu and Liu (2004a) pertaining to five different digital products. Each token in each sentence is labeled as described in section 2. In order to make robust comparisons and to be comparable with previous work, for each domain we create three random splits of the data with a train/development/test ratio of 3:1:1 (see Table 1). Since results may vary across random seeds (Dodge et al., 2020), we repeat each experiment using three different seeds and the final result is reported as the mean F1 score (and standard deviation) calculated over the three splits and the three seeds.

Domain	# Sentences	Train	Dev.	Test
(R)estaurant	5,841	4,381	1,460	1,460
(L)aptop	3,845	2,884	961	961
(D)evice	3,836	2,877	959	959

Table 1: Sentence statistics for each domain.

We adopt the HuggingFace (Wolf et al., 2019) implementation of BERT-base (uncased)[1] model as the basis for all experiments, and open-source our code.[2] We fine-tune the model with a learning rate of $5e^{-5}$, a batch size of 16 and a maximum sequence length of 64 tokens, for 10 epochs with an early stopping mechanism according to the development set. The dependency relations obtained by the Biaffine parser (Dozat and Manning, 2017) are generated in advance and are introduced to the model during the fine-tuning as well as during the development/test stages. Following prior work, only exact matches between the predicted aspect and opinion terms and the gold labels are counted as correct.

Results. Table 2 shows a comparison of our proposed model (**SA-EXAL**) with notable baseline models, across different domain transfers. The baselines include:

- **CrossCRF** (Jakob and Gurevych, 2010): A linear-chain CRF with hand-engineered features (e.g. POS tags and dependencies).
- **Hier-Joint** (Ding et al., 2017): An RNN with auxiliary labels derived from manually designed rules that are based on frequently observed syntactic relations between aspect and opinion terms.
- **RNCRF** (Wang et al., 2016): A joint recursive neural network and CRF for in-domain aspect and opinion terms extraction.
- **ARNN-GRU** (Wang and Pan, 2020): A dependency-tree-based recursive neural network with GRU which uses an auto-encoder in the auxiliary task to reduce label noise.
- **TRNN-GRU** (Wang and Pan, 2020): An extension of ARNN-GRU which integrates a conditional domain-adversarial network that takes both word features and syntactic head relations as input.
- **EXAL**: A baseline model that shares the same size and structure as the proposed model SA-EXAL (Section 3) but does not incorporate syntactic information.

Our proposed model (SA-EXAL) shows an advantage over EXAL which demonstrate that although it was shown that the pre-trained BERT model captures significant linguistic knowledge, informing it with explicit external dependency relations is effective for transferring knowledge across domains. Specifically, SA-EXAL outperforms EXAL in 10 out of 12 cases (underlined in the table), including 6.44%, 3.56% and 2.33% improvements for L $\rightarrow$ R (AS), R $\rightarrow$ L (AS) and R $\rightarrow$ D (AS), respectively. We also note that SA-EXAL outperforms the non pre-trained model baselines in 8 out of 12 cases.

[1] https://github.com/huggingface/transformers

[2] https://github.com/NervanaSystems/nlp-architect/tree/libert/nlp_architect/models/libert

	R → L		R → D		L → R		L → D		D → R		D → L	
Model	AS	OP	AS	OP	AS	OP	AS	OP	AS	OP	AS	OP
CrossCRF*	19.72 (1.82)	59.2 (1.34)	21.07 (0.44)	52.05 (1.67)	28.19 (0.58)	65.52 (0.89)	29.96 (1.69)	56.17 (1.49)	6.59 (0.49)	39.38 (3.06)	24.22 (2.54)	46.67 (2.43)
Hier-Joint*	33.66 (1.47)	- -	33.20 (0.52)	- -	48.10 (1.45)	- -	31.25 (0.49)	- -	47.97 (0.46)	- -	34.74 (2.27)	- -
RNCRF*	24.26 (3.97)	60.86 (3.35)	24.31 (2.57)	51.28 (1.78)	40.88 (2.09)	66.50 (1.48)	31.52 (1.40)	55.85 (1.09)	34.59 (1.34)	63.89 (1.59)	40.59 (0.80)	60.17 (1.20)
ARNN-GRU*	40.43 (0.96)	65.85 (1.50)	35.10 (0.62)	60.17 (0.75)	52.91 (1.82)	72.51 (1.03)	40.42 (0.70)	61.15 (0.60)	48.36 (1.14)	73.75 (1.76)	51.14 (1.68)	**71.18** (1.58)
TRNN-GRU*	40.15 (0.77)	65.63 (1.01)	37.33 (0.90)	60.32 (0.66)	53.78 (0.91)	73.40 (0.45)	41.19 (1.06)	60.20 (1.56)	51.17 (0.99)	**74.37** (1.03)	**51.66** (1.27)	68.79 (1.63)
EXAL	44.03 (2.11)	75.01 (1.13)	38.17 (0.79)	**63.59** (3.53)	48.23 (2.87)	79.57 (0.53)	41.60 (0.54)	**60.71** (5.49)	53.75 (1.24)	70.03 (2.46)	45.75 (1.54)	62.65 (2.51)
SA-EXAL	**47.59** (1.88)	**75.79** (1.02)	**40.50** (1.05)	63.33 (2.63)	**54.67** (2.02)	**80.05** (0.48)	**42.19** (0.54)	60.19 (3.79)	**54.54** (1.90)	71.57 (2.86)	47.72 (2.79)	63.98 (3.37)

Table 2: Comparison across different baselines in terms of average F1 scores (and standard variations in parentheses). *Results for non pre-trained baselines reported by (Wang and Pan, 2020). The best result for each dataset is highlighted in bold and the best result between EXAL and SA-EXAL is underlined.

5 Conclusion

We propose a method for incorporating external linguistic information into a self-attention mechanism coupled with the BERT model. We demonstrate that this model leverages both the intrinsic knowledge existing within the pre-trained model and the externally introduced syntactic information, to bridge the gap across domains.

Acknowledgements

The authors would like to thank Roi Reichart, Ido Dagan, Yael Klein, Intel AI Lab members and the anonymous reviewers for their valuable comments and feedback.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for crossdomain opinion target extraction. In *Association for the Advancement of Artificial Intelligence*, pages 3436–3442.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

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

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? In *arXiv*.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *American Association for Artificial Intelligence*.

Niklan Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045.

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, Lisbon, Portugal, September. Association for Computational Linguistics.

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. *CoRR*, abs/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, Dublin, Ireland, August. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37(1):9–27, March.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995, Long Beach, California, USA, 09–15 Jun. PMLR.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium, October-November. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1–11.

Wenya Wang and Sinno Jialin Pan. 2020. Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. *Computational Linguistics*, 45(4):705–736.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in NaturalLanguage Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 616–626.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.