Open Aspect Target Sentiment Classification with Natural Language Prompts

Ronald Seoh[*1] **Ian Birle**[*1] **Mrinal Tak**[*1] **Haw-Shiuan Chang**[*1]
Brian Pinette[2] **Alfred Hough**[2]
[1] University of Massachusetts Amherst [2] Lexalytics, Inc.
{bseoh, ibirle, mtak, hschang}@cs.umass.edu
{brian.pinette, al.hough}@lexalytics.com

Abstract

For many business applications, we often seek to analyze sentiments associated with any arbitrary aspects of commercial products, despite having a very limited amount of labels or even without any labels at all. However, existing aspect target sentiment classification (ATSC) models are not trainable if annotated datasets are not available. Even with labeled data, they fall short of reaching satisfactory performance. To address this, we propose simple approaches that better solve ATSC with natural language prompts, enabling the task under zero-shot cases and enhancing supervised settings, especially for few-shot cases. Under the few-shot setting for SemEval 2014 Task 4 laptop domain, our method of reformulating ATSC as an NLI task outperforms supervised SOTA approaches by up to 24.13 accuracy points and 33.14 macro F1 points. Moreover, we demonstrate that our prompts could handle *implicitly* stated aspects as well: our models reach about 77% accuracy on detecting sentiments for aspect categories (e.g., food), which do not necessarily appear within the text, even though we trained the models only with explicitly mentioned aspect terms (e.g., fajitas) from just 16 reviews — while the accuracy of the no-prompt baseline is only around 65%.

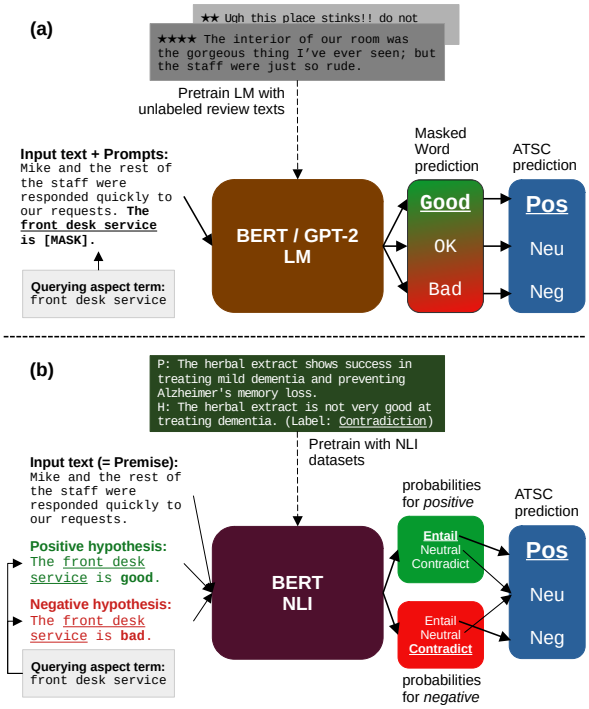

Figure 1: An overview of our prompt-based models for ATSC. (a) BERT / GPT-2 LM is pretrained on unlabeled review texts and we convert ATSC into a language modeling task during training and testing. (b) BERT NLI is pretrained on NLI datasets and we convert ATSC into two entailment tasks.

1 Introduction

Measuring targeted sentiments from text toward certain aspects or subtopics has immediate commercial value. For example, a hotel chain might want to base their business decisions on the proportion of customer reviews being positive toward their room cleanliness and front desk services. Manually reading through thousands of reviews would be prohibitively expensive, calling for automated solutions.

Adopting existing supervised models for aspect target sentiment classification (ATSC) (Pontiki et al., 2014) appears to be an obvious choice at first glance. However, the accuracy of these models in practice are often unsatisfactory due to the lack of labeled data from domains of interest. We could direct practitioners to annotate few domain-specific review texts, but ATSC models would need to generalize from those limited examples and return accurate predictions upon seeing any arbitrary aspects and sentiments. Even if we could get hold of more data, it would be difficult to collect enough labels to handle all aspect-sentiment cases that customers could ever query.

The way people would write customer reviews inspires our solutions to this label scarcity issue.

[*]Equal contribution.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6311–6322
November 7–11, 2021.

For example, consider two plausible sentences from hotel reviews: *Mike and the rest of the staff were very polite and responded quickly to our requests. The front desk service is excellent.* Supposing we only observe the first one, if an ATSC model could tell that the first *entails* the second, or the second *naturally follows* the first, we could determine that the customer feels positive about the front desk service.

Based on this intuition, we propose two ATSC models based on natural language prompts: our first method appends cloze question prompts (e.g., *The service is*) to the review text and predicts how likely it is to observe *good*, *bad*, and *ok* as the next word. The second method treats the review as a premise and the prompt sentence with the sentimental next word as a hypothesis, and predicts whether the review entails the prompt. With pretraining, these prompts give us natural ways of leveraging more abundant unlabeled reviews or large natural language inference (NLI) datasets to overcome the scarcity of labeled data. An overview of our methods is provided in Figure 1.

Most of previous ATSC efforts focus on fully supervised training, assuming enough in-domain labels on relevant aspects. Going beyond this typical setting, we also test our model on more "open" situations where reviews and aspects for testing are less similar to the training examples.

Experimental results show that our methods are not only capable of producing reasonable predictions without any labeled data (zero-shot) but also constantly outperform SOTA no-prompt baselines under supervised settings. Moreover, our prompt-based models are robust to *domain shifts*: after training with aspect-based sentiments in one domain, we observe that the models can accurately predict sentiments associated with aspect terms appearing in reviews from a completely unseen domain. Also, our models that are trained on explicitly mentioned aspect terms could generalize well to implied *categorical* aspects.[1]

2 Related Work

2.1 Aspect-based Sentiment Analysis

Hu and Liu (2004) is one of the first academic work to discuss the task of analyzing opinions targeted towards different aspects or topics within the text. Pontiki et al. (2014) starts the current line of research on aspect-based sentiment analysis (ABSA), with their benchmark datasets of customer reviews for restaurants and laptops.

To overcome small training dataset sizes, recent developments for ABSA involve some combination of unlabeled domain text pretraining and intermediate task finetuning. Xu et al. (2019) and Sun et al. (2019) use multi-task loss functions to finetune BERT using SQuAD question answering datasets that contain a range of domains and task-related knowledge, to offset the small size of SemEval 2014 ABSA datasets. Rietzler et al. (2020) provides detailed analysis on *cross-domain adaptation*, where they find that end task finetuning with the domain different from the evaluation domain still achieves performance comparable to the SOTA results using in-domain labels.

2.2 Natural Language Prompts

There has been a number of recent papers on using prompts — additional sentences appended to the original input text — to direct language models to perform different tasks, exploiting the knowledge they have acquired during their original pretraining. One of the earliest examples of such efforts is Radford et al. (2019), where they measured their GPT-2 model's performance on downstream tasks by feeding in task descriptions as prompts, without any finetuning at all. Since then, a number of previous work has leveraged prompts for the tasks such as question answering (Lewis et al., 2019) and commonsense reasoning (Shwartz et al., 2020). We also note that some previous work on prompt-based learning methods have included sentence-level sentiment classification, which measures sentiment from the entire input text, as part of their evaluation (Shin et al., 2020; Gao et al., 2021).

Many recent works, including ours, follow the format of *cloze questions* to design prompts as first suggested by Schick and Schütze (2020). We design prompts to include masked tokens that need to be filled in, and the predictions for the masks serve as the outputs for the original task.

3 Methods

To maximally leverage the large data resources of unlabeled review texts and NLI datasets, we propose two ways of reformulating ATSC: the first converts ATSC into next/masked word prediction; the second transforms the task into NLI entailment

[1] All the program codes used to produce results presented in this paper are available at https://link.iamblogger.net/atscprompts.

predictions.

3.1 ATSC as Language Modeling

In order to elicit abilities to perform ATSC from language models (LMs), it was essential that prompt sentences should be similar to what one would typically write to express their sentiment. Hence, we came up with the following set[2] of cloze question prompts (Schick and Schütze, 2020) that are aspect dependent, and appended them to input texts:

- `I felt the {aspect} was [MASK].`
- `The {aspect} made me feel [MASK].`
- `The {aspect} is [MASK].`

where `{aspect}` is the placeholder for the querying aspect term, and `[MASK]` represents the masked word for BERT (Devlin et al., 2019) and the next word for GPT-2 (Radford et al., 2019). Then, we let the probability of predicting positive, neutral, and negative sentiment be proportional to the probability of predicting `good`, `ok`, and `bad`, respectively.

Publicly available pretrained weights for BERT and GPT-2 have already been trained on large general corpora, but they might not include enough sentences from our testing domains, such as laptop reviews. To produce more accurate predictions on `[MASK]`, we further pretrain LMs with the in-domain review texts. For BERT, we modify the original random masking scheme to mask only adjectives, nouns, and proper nouns because the words are more likely to indicate the sentiments of the sentence. For GPT-2, we use the original causal LM (CLM) objective. To measure the effectiveness of the prompts, our baselines without prompts also receive identical pretraining.[3]

When the labeled data for ATSC is available, we convert the training labels to `good`, `ok`, and `bad` and finetune all the parameters of LMs, including the encoders and embeddings of words in the prompts. During the training and testing, other candidates for `[MASK]` are ignored.

3.2 ATSC as NLI

We first set the input review text as a premise. We predict the scores for positive, negative, and neutral sentiment using the entailment probabilities from a NLI model as follows: we create positive and negative hypotheses by populating prompts with corresponding label words (e.g., `"The {aspect} is good; The {aspect} is bad."`). We get the scores for positive and negative sentiment by obtaining entailment probabilities with each of the hypotheses; For the neutral class, we average neutral probabilities (from NLI) for the two hypotheses. Our method enables zero-shot ATSC, which previously had not considered by previous efforts that leveraged NLI for sentiment analysis and text classification tasks (Yin et al., 2019; Sun et al., 2019; Wang et al., 2021).

We use the BERT-base model pretrained on the MNLI dataset (Williams et al., 2018), and none of the unlabeled review texts are utilized for pretraining. We apply a softmax layer on top of the logits to normalize the prediction scores of the three classes into probabilities, in order to finetune models with cross-entropy loss when labeled data is available.

4 Experiments

In the experiments, we test the generalization capability of our methods on more real world-like conditions where there are far fewer training examples similar to the testing examples. Using ATSC datasets from SemEval 2014 Task 4 Subtask 2 (Pontiki et al., 2014), we evaluate our models on the full spectrum of in-domain training data sizes covering the zero-shot and full-shot (i.e., fully supervised) cases. Similar to the settings of Scao and Rush (2021), we train our models with randomly re-sampled training sets of sizes {Zero, 16, 64, 256, 1024, Full}.

Furthermore, we conduct *cross-domain* evaluation where we train the models on restaurant reviews and test on laptop reviews, and vice versa. Finally, we train the models on ATSC and test them on aspect *category* sentiment classification (ACSC), another ABSA variant, to evaluate the robustness to an unseen querying aspect distribution. Unlike aspect terms, these categories such as *food, service, price, and ambience* usually do not explicitly appear within the text, but *implicitly* stated through aspect terms or overall context.[4]

[2] In our experiments, all three prompts perform similarly, especially in few-shot cases; this suggests that the performance of our models are not overly sensitive to the wording of properly chosen prompts. We present the performance comparison between our prompts in Appendix C.

[3] We compare the results between original weights and our further pretrained ones in Appendix F.

[4] Please refer to Appendix A for ATSC/ACSC dataset statistics and preprocessing.

Model	Number of Training Examples: Zero Acc	Zero MF1	16 Acc	16 MF1	64 Acc	64 MF1	256 Acc	256 MF1	1024 Acc	1024 MF1	Full (1850) Acc	Full (1850) MF1
BERT-ADA	–	–	–	–	–	–	–	–	–	–	79.19†	74.18†
BERT [CLS]	–	–	48.75	34.92	60.63	49.43	72.35	64.31	76.87	71.22	80.06	75.08
BERT NSP	–	–	48.24	31.35	60.91	49.27	72.38	64.64	76.77	71.12	80.25	75.46
BERT LM	**63.58**	46.17	69.05	58.60	72.80	65.54	**76.59**	70.65	**79.30**	74.80	**81.10**	76.83
	–	–	+20.30*	+23.68*	+11.89*	+16.11*	**+4.21***	+6.01*	**+2.43***	+3.58*	**+0.85***	+1.37*
GPT-2 LM	60.45	39.59	68.94	56.71	71.54	63.69	76.48	70.89	79.02	**74.88**	80.73	**77.13**
	–	–	+20.19*	+21.79*	+10.63*	+14.26*	+4.10*	+6.25*	+2.15*	+3.66*	+0.48	**+1.67***
BERT NLI	58.93	**54.91**	**72.88**	**68.06**	**74.95**	**70.84**	76.22	**71.65**	77.42	73.52	77.58	73.18
	–	–	**+24.13***	**+33.14***	**+14.04***	**+21.41***	+3.84*	**+7.01***	+0.55	+2.30	-2.67	-2.28

(a) Laptops

	Zero Acc	Zero MF1	16 Acc	16 MF1	64 Acc	64 MF1	256 Acc	256 MF1	1024 Acc	1024 MF1	Full (3602) Acc	Full (3602) MF1
BERT-ADA	–	–	–	–	–	–	–	–	–	–	87.14†	80.05†
BERT [CLS]	–	–	59.89	34.50	73.00	50.79	79.45	64.70	83.48	73.62	86.77	79.33
BERT NSP	–	–	61.05	32.46	74.73	53.00	79.34	65.51	83.61	74.15	87.09	79.98
BERT LM	70.86	48.17	71.99	56.65	77.79	63.30	81.10	69.27	85.12	76.60	**87.50**	**80.78**
	–	–	+10.94*	22.15*	+3.06*	+10.30*	+1.65*	+3.76*	+1.51*	+2.45*	**+0.41**	**+0.80**
GPT-2 LM	**71.40**	45.53	**75.41**	60.06	79.30	65.49	**82.27**	71.62	**85.28**	**77.38**	86.99	80.02
	–	–	**+14.36***	+25.56*	+4.57*	+12.49*	**+2.82***	+6.11*	**+1.67***	**+3.23***	-0.1	+0.04
BERT NLI	61.79	**57.93**	74.74	**65.58**	**79.33**	**69.44**	81.24	**71.94**	83.07	74.52	85.07	77.53
	–	–	+13.69*	**+31.08***	**+4.60***	**+16.44***	+1.79*	**+6.43***	-0.54	+0.37	-2.02	-2.45

(b) Restaurants

Table 1: Results of our methods and baselines. Acc and MF1 refer to accuracy and macro F1, respectively. We use five random seeds for each of the prompts and baselines, and average their scores. We averaged the performance of our models across all three prompts. Please see Appendix C for performance comparison between the prompts. Boldfaces indicate the best performance given the same number of labels, and the best baseline scores are underlined. † BERT-ADA results are taken directly from Rietzler et al. (2020). * indicates an increase over the baseline with significance level .05 using a two mean z-test.

4.1 Baselines

We compared our prompt-based methods with two common strategies of utilizing BERT for classification tasks: 1) the last hidden state of `[CLS]` token (BERT [CLS]), and 2) the NSP head of BERT (BERT NSP). We note that the architecture of BERT NSP is equivalent to BERT-ADA (Rietzler et al., 2020), currently the top-performing BERT-based model for ATSC which we show their reported full-shot performance for reference.

4.2 Results

Prompts constantly outperform the no-prompt baselines. We can see in Table 1 that for both target domains, our prompt-based BERT models outperform the no-prompt baselines in all cases. Especially for few-shots, we achieve larger performance gains as fewer labels are available. We note that BERT NLI does particularly well in 16 to 256 shots for laptops, with noticeably higher accuracy and macro F1 (MF1) than other prompt models.[5]

We emphasize again that our NLI models are only trained on the MNLI dataset, which makes them particularly preferable when in-domain text (shopping reviews) is not readily available.

Lastly, we observe that our methods achieve good performances in the zero-shot cases, significantly outperforming the baselines that are trained on 16 samples, further showing its practicality.

Prompts can utilize cross-domain data more effectively. As shown in Table 2, the prompt models with 16-shot cross-domain training achieve better performance than both in- and out-domain BERT NSP. It is also interesting to note that cross-domain have even exceeded in-domain for the restaurants domain. This suggests that our methods might have the potential to be particularly more

[5]With the help of more abundant neutral examples in MNLI, Appendix G suggests that BERT NLI is particularly better at detecting neutral sentiment, subsequently leading to better MF1 scores.

Model	In/Cross	16		Full	
		Restaurants	Laptops	Restaurants	Laptops
BERT NSP	In	61.05	48.24	87.09	80.25
	Cross	49.55	47.46	81.29	78.56
BERT LM	In	71.99	69.05	87.50	81.10
	Cross	75.21	68.17	81.27	79.03
BERT NLI	In	74.74	72.88	85.07	77.58
	Cross	77.45	70.43	80.35	76.61

Table 2: Accuracies of BERT NSP, LM, and NLI trained with in-domain and cross-domain data.

Model	16		Full	
	Acc	MF1	Acc	MF1
BERT NSP	64.73	33.22	82.45	70.91
BERT LM	76.67	56.77	84.31	74.14
BERT NLI	66.92	58.18	67.42	59.24

Table 3: Performance on ACSC without any extra training. Refer to Appendix E for the results with other training set sizes.

adaptable to arbitrary domains under low-resource settings, which we plan to explore further in future research.

Prompts can better recognize *implicit* aspects. We can see from Table 3 that the BERT LM model trained with merely 16 examples achieves about 77% accuracy on ACSC, while it was never trained in terms of aspect categories at all. BERT NSP, the no-prompt baseline, achieves around 65%. This result suggests that our prompt-based models have also acquired some abilities to recognize aspects that are implied or worded differently from the querying aspect term. Such abilities could also make our prompt models more desirable for potential real-life applications. We note that BERT NLI performs rather poorly, particularly under full-shot. As it hadn't seen in-domain texts during pre-training, we suspect that it cannot fully recognize related domain-specific words.

5 Conclusion and Future Work

In this paper, we examined our prompt-based ATSC models leveraging LM and NLI under zero-shot, few-shot, and full supervised settings. We observe a significant amount of improvements over the no-prompt baselines in nearly all configurations we have tested. In particular, we find that our NLI model performs well with lower amounts of training data, while the BERT LM model does better when more labels are available. In addition, we have seen that it could effectively utilize cross-domain labels and recognize implicit aspects, suggesting that it would potentially be more applicable in real-life scenarios.

For future work, one direction is to adapt our aspect-dependent prompts to the models that jointly perform aspect term extraction and sentiment classification, such as Luo et al. (2020). Secondly, we could explore potential ways of combining ATSC, masked/next word prediction, and NLI into a unified task in order to take the full advantage of both our unlabeled text and NLI pretraining. Lastly, it would be an interesting analysis to determine whether there are any strong linguistic patterns among correct or incorrect predictions that each of our models make — such findings could allow us to have more detailed insights into the potential behaviors of our prompt-based models.

Acknowledgments

We thank our anonymous reviewers for providing valuable feedback. This work was supported by the Center for Data Science at the University of Massachusetts Amherst, under the Center's industry mentorship program. We would like to express our gratitude to Professor Mohit Iyyer and Professor Andrew McCallum for suggesting constructive discussions throughout the course of our project. We also would like to thank Paul Barba of Lexalytics for setting the project's initial research directions. Last but not least, we appreciate all the efforts Xiang Lorraine Li and Rico Angell have put in to manage the program participants and provide various practical assistance.

This work was produced in part using high performance computing equipment obtained under a grant from the Collaborative R&D Fund, managed by the Massachusetts Technology Collaborative.

References

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. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830, Online. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

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

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

Huaishao Luo, Lei Ji, Tianrui Li, Daxin Jiang, and Nan Duan. 2020. GRACE: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 54–64, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

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

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few-shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Hu Xu, Bing Liu, Lei Shu, and Philip 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A Dataset Information

A.1 SemEval 2014 Task 4 dataset

Aspect Target Sentiment Classification (ATSC, Subtask 2) The dataset released by Pontiki et al. (2014) is one of the most popular benchmark datasets for ATSC used in the literature, which contains English review sentences from two target domains, laptops, and restaurants. We measure the performance of our models and baselines on the test splits of each domain.

Labels for the sentiments are limited to positive, negative, neutral, and conflict. neutral refers to the case where the opinion towards the aspect is neither positive nor negative, and conflict is for both positive and negative sentiments being expressed for *one* aspect.

To make our work comparable with previous efforts (Xu et al., 2019; Rietzler et al., 2020), we use the following preprocessing procedure:

1. Reviews with conflict labels were removed. They have been usually ignored due to having a very small number of examples.

2. Multiple aspect-sentiment labels within one text piece were split up into different data points.

Dataset statistics after preprocessing are provided in Table 4.

Class	Restaurant		Laptop	
	Train	Test	Train	Test
Positive	2164	728	987	341
Negative	645	196	866	128
Neutral	496	196	460	169
All	3602	1120	1850	638

Table 4: SemEval 2014 dataset statistics after preprocessing.

Aspect Category Sentiment Classification (ACSC, Subtask 4) For this task, the labeled data is available only for the restaurant domain. While the class labels are the same as ATSC, this task has predefined aspect categories: food, price, service, ambience, anecdotes/miscellaneous. We only use the test split with 973 examples, containing 657 positives, 222 negatives, and 94 neutrals. The train split for ACSC is never used in any manner throughout our experiments.

A.2 LM Pretraining Corpora

We note the following sources of unlabeled review texts to further pretrain BERT and GPT-2 language models for two target domains of SemEval 2014 Task 4:

1. **Amazon Review Data** (Ni et al., 2019) is the collection of customer reviews extracted from the online shopping website Amazon. We used 20,994,353 reviews written for the products from the electronics category. The LMs pretrained with this collection are used to target the ATSC laptop domain.

2. **Yelp Open Dataset**[6] consists of over 8 million business reviews. We extracted the reviews associated with restaurants (2,152,007 reviews). The LMs pretrained with this collection are used to target the ATSC restaurants domain.

B Training and Testing Settings

Our training and testing settings are summarized in Figure 2.

All the program codes used to produce results presented in this paper are available at `https://link.iamblogger.net/atscprompts`.

Publicly available BERT MLM, GPT-2 CLM, and BERT NLI weights Following the common practice in recent NLP transfer learning literature, we use publicly available weights pretrained on large unlabeled corpora and task datasets for further training. For BERT MLM and GPT-2 CLM, we use the weights obtained from the `transformers` library (Wolf et al., 2020). For BERT NLI, we use the weights released by Morris et al. (2020), which were trained on the MNLI dataset (Williams et al., 2018).

BERT / GPT-2 main layer finetuning Unlike more usual ways of using prompts where the main layers of language models are left frozen and do not get to see any updates, we simply leave them open for further finetuning. While this technically leads to more amount of computation, we anticipated that the cost would be negligible given that the amount of labeled data for the task is fairly small, especially for few-shot learning cases we are particularly interested in.

[6] `https://www.yelp.com/dataset`

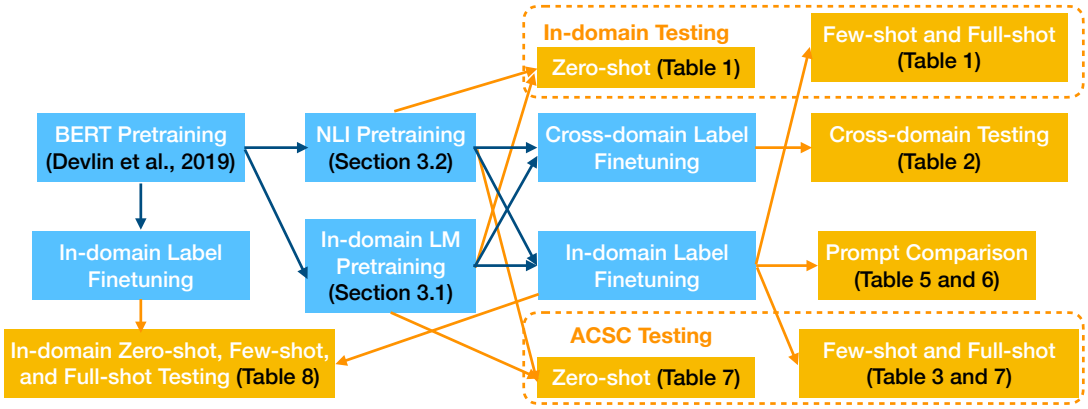

Figure 2: The flow chart of our methods and experiments. The blue blocks represent the training steps and the orange blocks represent the testing steps.

Train-test split and training hyperparameters For the full-shot training, we use the entirety of the train split of the SemEval dataset following Rietzler et al. (2020).

For ATSC training, we train with the SemEval datasets for 20 epochs. Following the recommendations regarding BERT finetuning made by Mosbach et al. (2020), we finetune all our prompt-based models and no-prompt baselines until the minimum training loss reaches near zero to achieve stable task performance, where the minimum losses are around 1e-07 and 1e-06.

For each ATSC model, we train them 5 times with different random seeds. Using different random seeds changes the data loading order, and the subset of training examples chosen for few-shot settings.

Hardwares and Softwares Used For each ATSC model, we used one NVIDIA TITAN X GPU for training. The version 4.3.3 of `transformers` library (Wolf et al., 2020) is used with `pytorch` version 1.7.1. We also implemented all the loading scripts for our datasets to be compatible with the version 1.2.1 of the Huggingface `datasets` library[7]. We have used the `spacy` library (Honnibal and Montani, 2017) for POS tagging, and `pytokenizations`[8] for tokenizer alignment.

Prompt	Accuracy (Std. Error)	Macro F1 (Std. Error)
BERT NSP (No prompt)	48.24 (0.0283)	31.35 (0.0198)
"I felt the {aspect} was [MASK]."	69.06 (0.0060)	59.71 (0.0214)
"The {aspect} made me feel [MASK]."	68.15 (0.0069)	56.59 (0.0205)
"The {aspect} is [MASK]."	69.94 (0.0061)	59.51 (0.0179)

(a) Laptops

Prompt	Accuracy (Std. Error)	Macro F1 (Std. Error)
BERT NSP (No prompt)	61.05 (0.0238)	32.46 (0.0374)
"I felt the {aspect} was [MASK]."	73.59 (0.0247)	59.03 (0.0168)
"The {aspect} made me feel [MASK]."	69.38 (0.0223)	51.65 (0.0114)
"The {aspect} is [MASK]."	73.02 (0.0209)	59.25 (0.0152)

(b) Restaurants

Table 5: Comparing different prompts on 16-shot training of our prompt models and baselines.

C Comparing Performance of Different Prompts

While Table 1 shows the scores averaged over different prompts, the performances are very similar across different manual prompts we have chosen, as seen during the full-shot training for the laptops domain in Table 5. We observed similar trend in the restaurant domain, and few-shot scenarios we

[7] https://huggingface.co/docs/datasets/master/

[8] https://github.com/tamuhey/tokenizations

have tested. This suggests that practically reasonable choices of prompts could still achieve good ATSC performance.

D Importance of Aspect Dependent Prompts

Prompt	Accuracy	Macro F1
"I felt the {aspect} was [good/bad]."	77.77	73.56
"I felt the things were [good/bad]."	74.06	70.69
	-3.71	-2.88

(a) Laptops

Prompt	Accuracy	Macro F1
"The aspect is [good/bad]."	85.93	79.18
"The things are [good/bad]."	74.80	64.95
	-11.13	-14.22

(b) Restaurants

Table 6: Performance changes with aspect terms removed from the best performing prompt for our NLI model.

We performed small sanity check experiments where we confirm that given aspect terms in the prompts are actually being utilized to produce correct predictions. We take the best performing prompts for our NLI model, and replace all aspect terms with `things` for all test examples in the full-shot setting, regardless of actual aspect terms. Then, the exact prompt wording would become global for all inputs. Table 6 shows that this brings significant drops in performance for both test domains, showing that the prompt needs to be aspect dependent to produce accurate predictions.

E Detailed ACSC Results

In Table 7, we can see that BERT NLI is generally performing worse than BERT LM and BERT NSP in most cases. Unlike BERT LM, its performance appears to be stagnating even with more ATSC training examples. This trend is very different from what we observed in the main ATSC results in Table 1, where BERT NLI maintained high performance advantages in few-shot cases. The most probable cause for BERT NLI's worse performance is that it cannot fully comprehend the entailment relationships expressed with domain-specific vocabularies - due to not having done in-domain text pretraining, it seems quite likely for BERT NLI that it cannot recognize the facts such as `fajita` being a sort of *food*.

F Effectiveness of In-domain LM Pretraining

In Table 8, we show both BERT NSP and BERT LM results with the original pretrained weights ("Original") the weights further trained with domain-specific review texts ("Amazon, Yelp"). We could see that BERT LM with the original weights performs better than BERT NSP with domain-specific further pretraining in few shot settings. As previously suggested in Scao and Rush (2021), it appears that just using prompts alone does bring the positive benefits of alleviating the labeled data scarcity. We also note that the gap between the original weights and the further pretrained ones is relatively small when more number of labeled examples becomes available for training.

G Further Error Analysis

Model	Restaurants			Laptops		
	Pos	Neg	Neu	Pos	Neg	Neu
BERT NSP	75.50	14.47	7.40	63.65	22.76	7.64
BERT LM	86.34	57.65	25.95	83.51	60.83	31.46
GPT-2 LM	87.95	65.68	26.54	82.85	66.83	20.47
BERT NLI	80.86	66.30	51.55	83.60	69.60	50.98

Table 10: F1 scores for each class achieved by the baseline and our models with 16 examples.

In Table 10, we show F1 scores for each classes. Over the baseline BERT NSP, all our prompt models show large improvements. Particularly with BERT NLI, we see that F1 for the neutral class has greatly improved, doing better than both BERT NSP and BERT LM. It appears that BERT NLI is particularly better than BERT LM at detecting neutral and negative examples, quite possibly because the MNLI dataset contains many examples with neutral and negative labels.

In Table 9, we present a few notable examples from test data that one or more of our prompt-based models had predicted correctly while the baseline did not. R1 shows an example, which our model correctly classifies as neutral while the no-prompt baseline wrongly predicts positive. R2 shows an example where our models are able to make the correct prediction despite having multiple aspects within one sentence. We found R3 quite interesting, where there are no explicit terms to express negative sentiment, intuitively making it difficult for the model to detect sentiment; Yet only the NLI model is able to make the correct prediction. For

Model	Number of ATSC Examples											
	Zero		16		64		256		1024		Full	
	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1
BERT NSP	–	–	64.73	33.22	81.05	56.77	84.60	70.88	85.28	74.52	82.45	70.91
BERT LM	76.40	50.11	76.67	56.77	82.75	64.26	84.70	68.31	86.28	74.80	84.31	74.14
	–	–	+11.94	+23.55	+1.70	+7.49	+0.1	-2.57	+1.00	+0.28	+1.86	+3.23
BERT NLI	44.36	40.77	66.92	58.18	73.41	63.67	69.52	60.66	70.81	61.61	67.42	59.24
	–	–	+2.19	+24.96	-7.64	+6.90	-15.08	-10.22	-14.47	-12.91	-15.03	-11.66

Table 7: Performance on ACSC test data with our ATSC prompt models and baselines. We use 5 random seeds for each of the prompts and baselines, and average their scores.

Model	Pretraining Corpora	Number of Training Examples											
		Zero		16		64		256		1024		Full	
		Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1
BERT NSP	Original	–	–	45.74	30.25	50.88	36.81	69.69	63.21	76.14	70.64	77.96	73.24
	Original + Amazon	–	–	48.24	31.35	60.91	49.27	72.38	64.64	76.77	71.12	80.25	75.46
				+2.5	+1.10	+10.03	+12.46	+2.69	+1.43	+0.63	+0.48	+2.29	+2.22
BERT LM	Original	59.20	38.42	65.25	55.82	70.54	63.30	73.33	66.86	76.67	71.73	77.61	73.06
	Original + Amazon	63.58	46.17	69.05	58.60	72.80	65.54	76.59	70.65	79.30	74.80	81.10	76.83
		+4.38	+7.75	+3.80	+2.78	+2.26	+2.24	+3.26	+3.79	+2.63	+3.07	+3.49	+3.77

(a) Laptops

Model	Pretraining Corpora	Number of Training Examples											
		Zero		16		64		256		1024		Full	
		Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1
BERT NSP	Original	–	–	55.00	34.09	63.36	37.26	74.39	58.16	80.57	70.09	84.77	76.93
	Original + Yelp	–	–	61.05	32.46	74.73	53.00	79.34	65.51	83.61	74.15	87.09	79.98
				+6.05	-1.63	+11.37	+15.74	+4.95	7.35	+3.04	+4.06	+2.32	+3.05
BERT LM	Original	68.04	36.44	65.73	51.93	74.88	58.21	77.24	63.76	81.82	72.21	84.26	75.90
	Original + Yelp	70.86	48.17	71.99	56.65	77.79	63.30	81.10	69.27	85.12	76.60	87.50	80.78
		+2.82	+11.73	+6.26	+4.72	+2.91	+5.09	+3.86	+5.51	+3.30	+4.39	+3.24	+4.88

(b) Restaurants

Table 8: Comparing our prompt model performance between the original pretrained weights ("Original") and the weights further trained with domain-specific review texts ("Amazon, Yelp").

Type	Review	Truth	Baseline	NLI	MLM	GPT-2
R1	the good place to hang out during the day after shopping or to grab a simple soup or classic french dish over a *glass of wine*.	Neu	Pos	Neu	Neu	Neu
R2	My friend had a *burger* and I had these wonderful blueberry pancakes.	Neu	Pos	Neu	Neu	Neu
R3	The *sushi* is cut in blocks bigger than my cell phone.	Neg	Neu	Neg	Neu	Neu
R4	The absolute worst service I've ever experienced and the food was below average (when they actually gave people the *meals* they ordered).	Neu	Neu	Neg	Neg	Neg
R5	*Food* was decent, but not great.	Pos	Neu	Neu	Neu	Neu

Table 9: Analysis of various scenarios where our models and baselines fail. Aspect words are underlined, predictions are highlighted in green (correct) or red (incorrect).

future studies, it would be an interesting direction to perform further statistical analysis on whether phenomena we see here generally hold for prompt-based models and strong linguistic patterns emerge among them.

We also note some examples that our models were not able to classify correctly. For R4, all the models got wrong except for the baseline. This example seems particularly challenging, as there is another sentence showing negative sentiment about a very similar aspect (“the food was below average") R5 shows an example where the baseline and all our models fail — we find the true label here somewhat questionable, as intuitively the reviewer indeed appears to be neutral about food and not particularly positive.