EMNLP 2021 Presentation
7-11 November 2021

UMassAmherst
LEXALYTICS

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang
Brian Pinette, Alfred Hough

Open Aspect Target Sentiment Classification with Natural Language Prompts

Introduction

Aspect Target Sentiment Classification (ATSC)

- Determine whether the sentiment **associated with the given term** is positive, negative, or neutral.

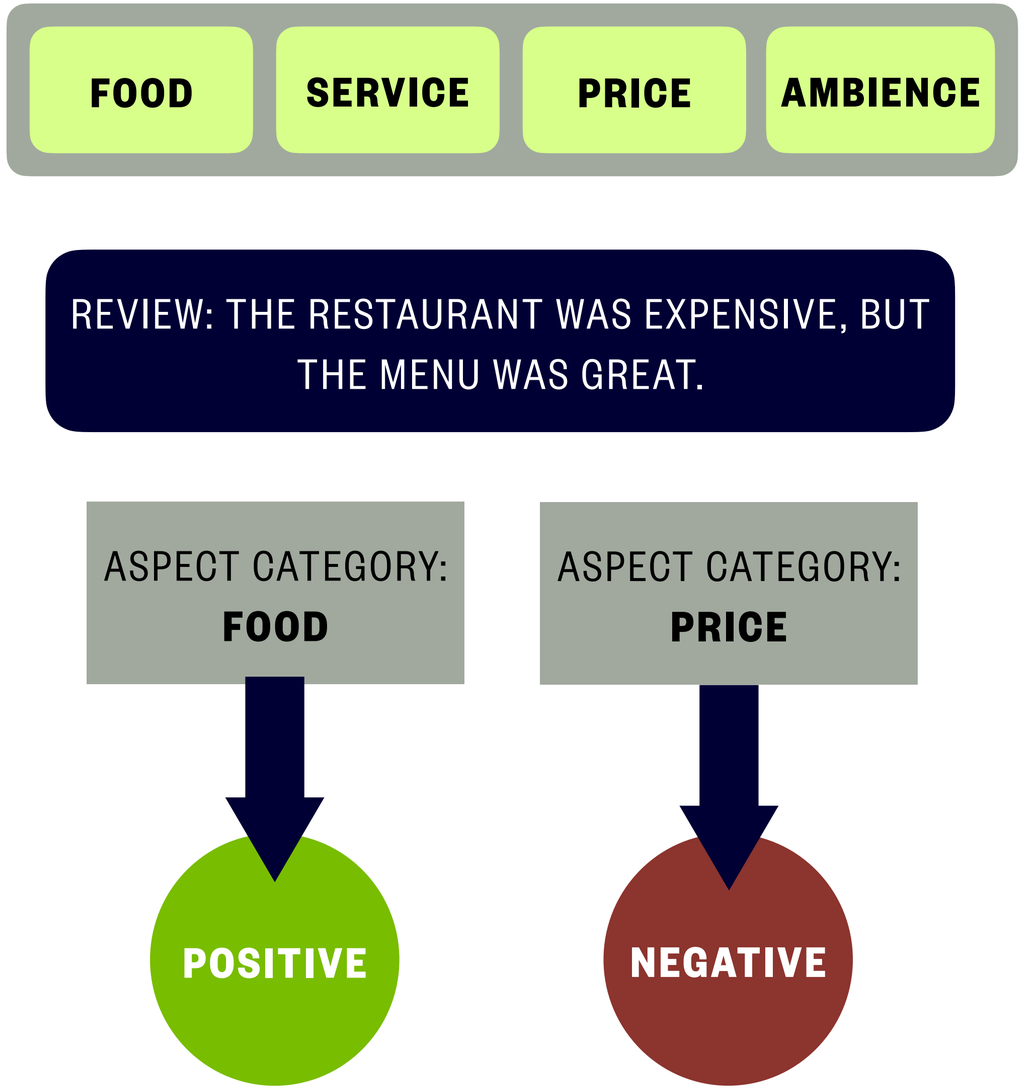
REVIEW: I HATED THEIR **FAJITAS**, BUT THEIR **SALADS** WERE GREAT.

ASPECT TERM: **SALADS** → **POSITIVE**

ASPECT TERM: **FAJITAS** → **NEGATIVE**

Pontiki et al. (2014), Semeval-2014 task 4: Aspect Based Sentiment Analysis.

Aspect Category Sentiment Classification (ACSC)



pre-defined aspect categories.

Aspect categories are not directly mentioned, but **<u>implicitly stated</u>**.

queries chosen from the pre-defined set.

Motivation:
Making ATSC possible under more realistic scenarios

Enable ATSC even when no training examples are available at all.

Exploit all labeled examples if available,
Both in-domain and cross-domain.

Handle unobserved variations in aspect-sentiment expressions.

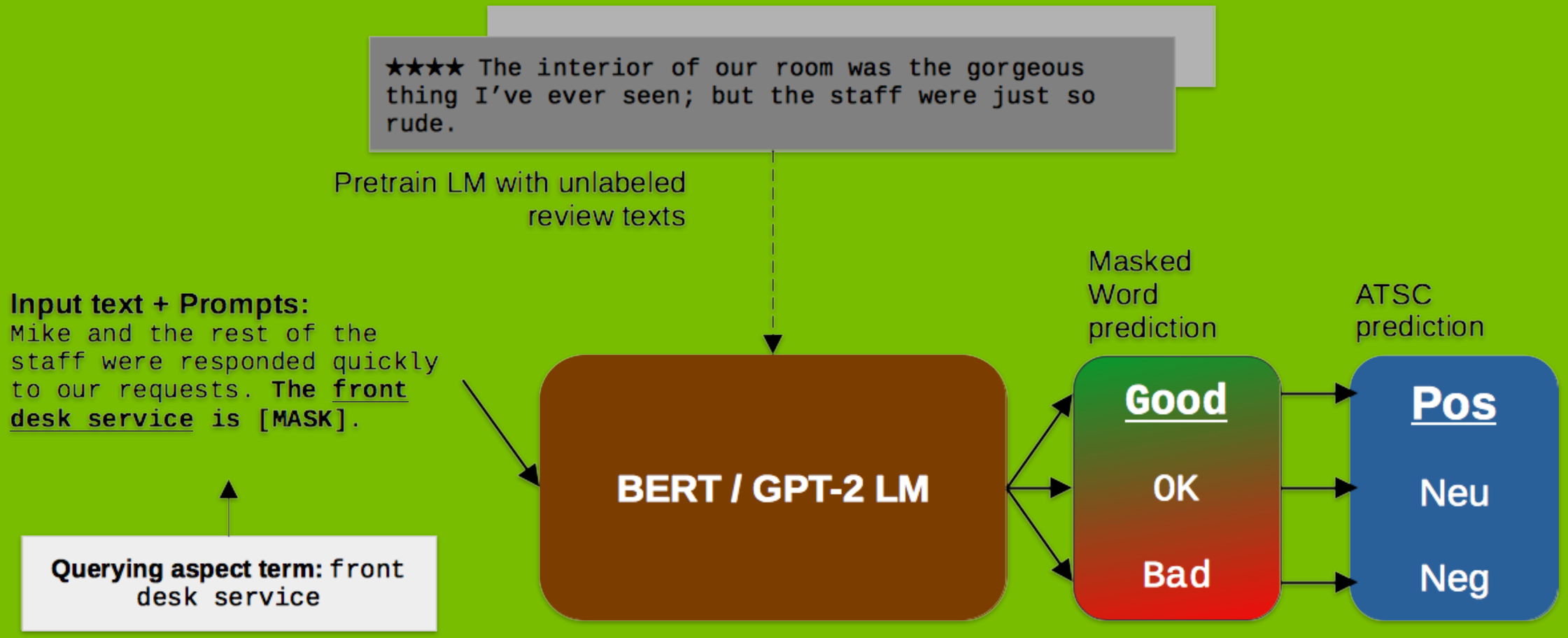
What if we could design an model that could...

Mike and other staff were very nice and polite. The front desk service is excellent.

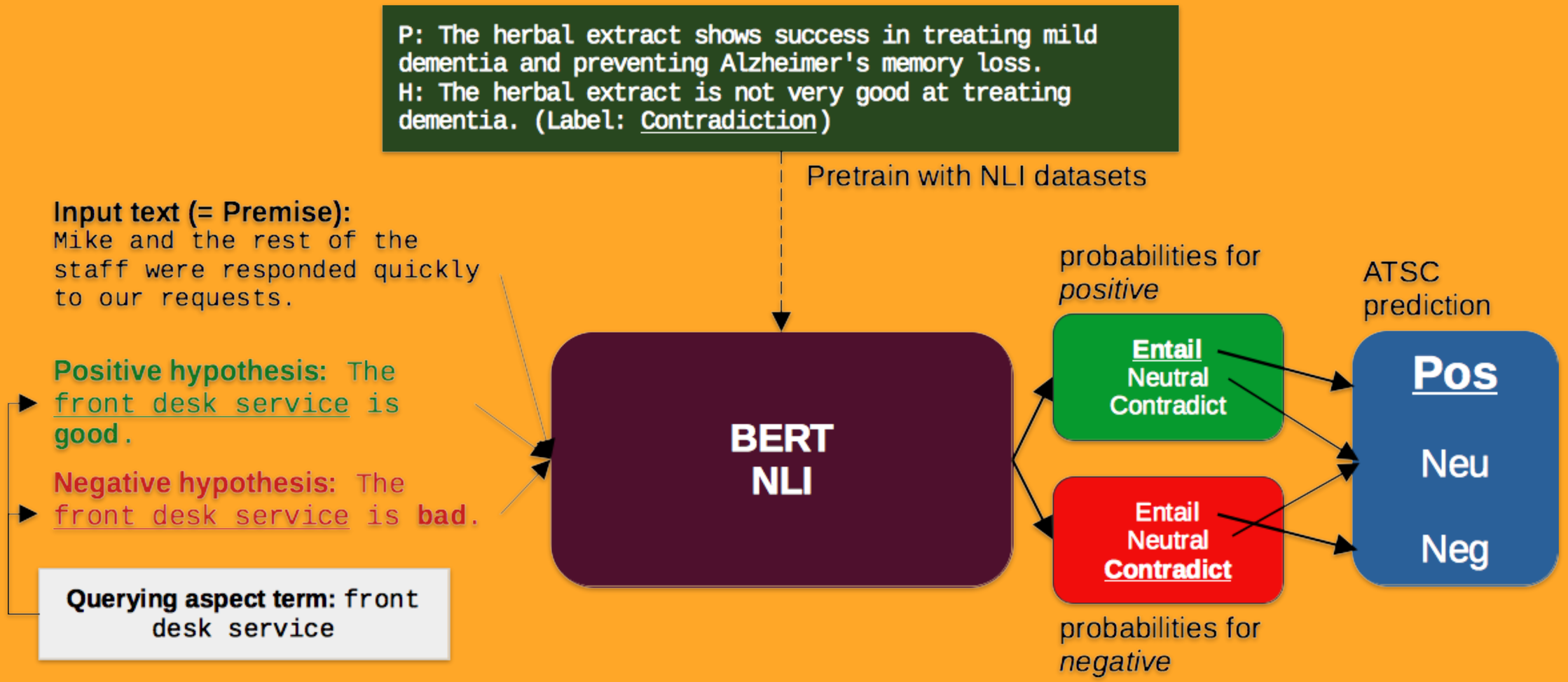
1) Predict whether the second sentence **naturally follows** the first?

2) Predict whether the first **entails** the second?

ATSC as a LM task



ATSC as a NLI task



Solving NLP tasks with language model (LM) prompting

Popularized by GPT-2/3;
We follow the recent trend of cloze question style prompts introduced by Schick and Schutze (2020).

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

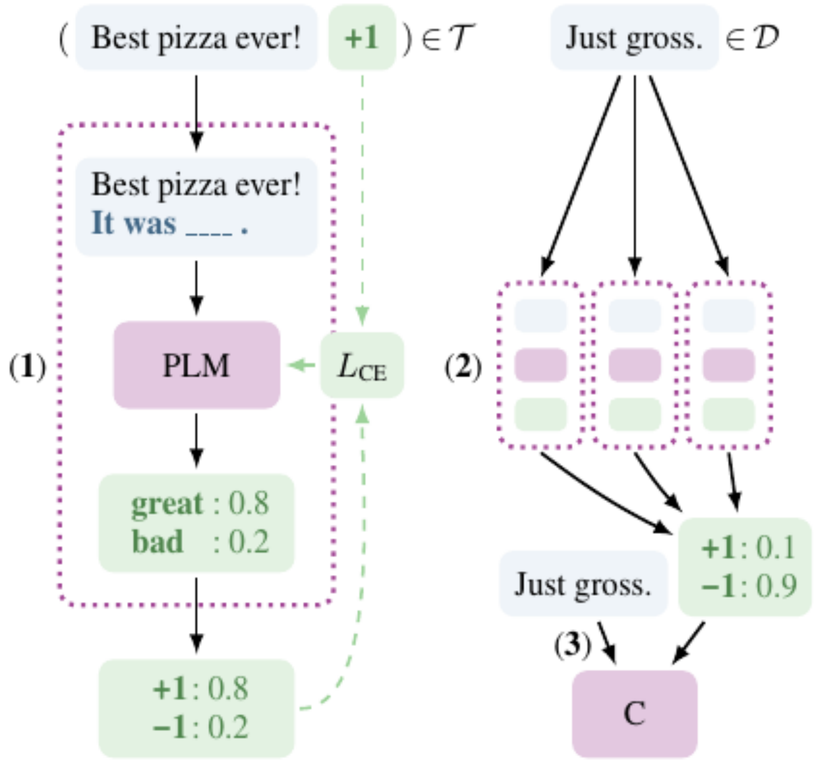
```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

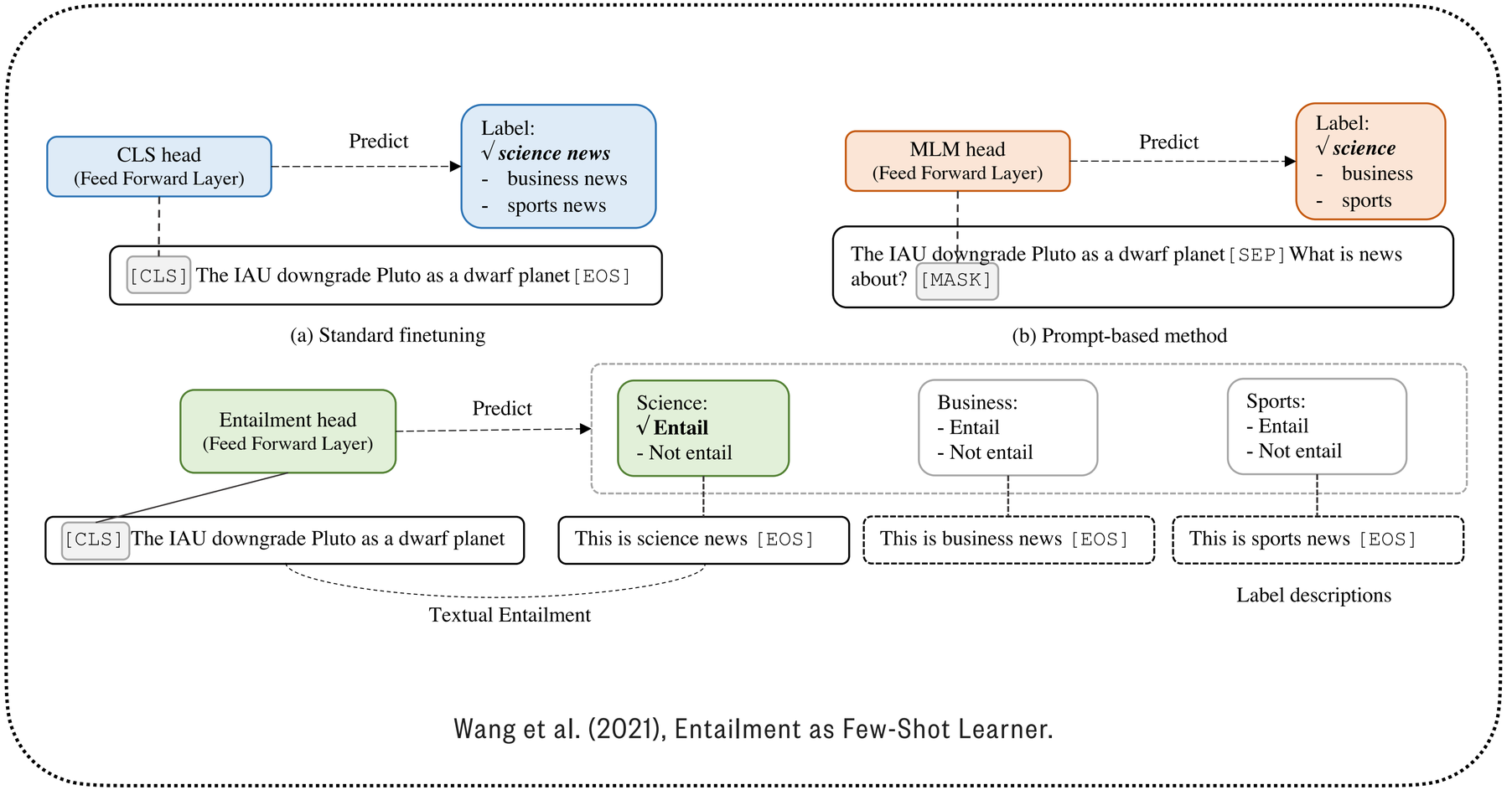
Brown et al. (2020),
Language models are Few-Shot Learners.



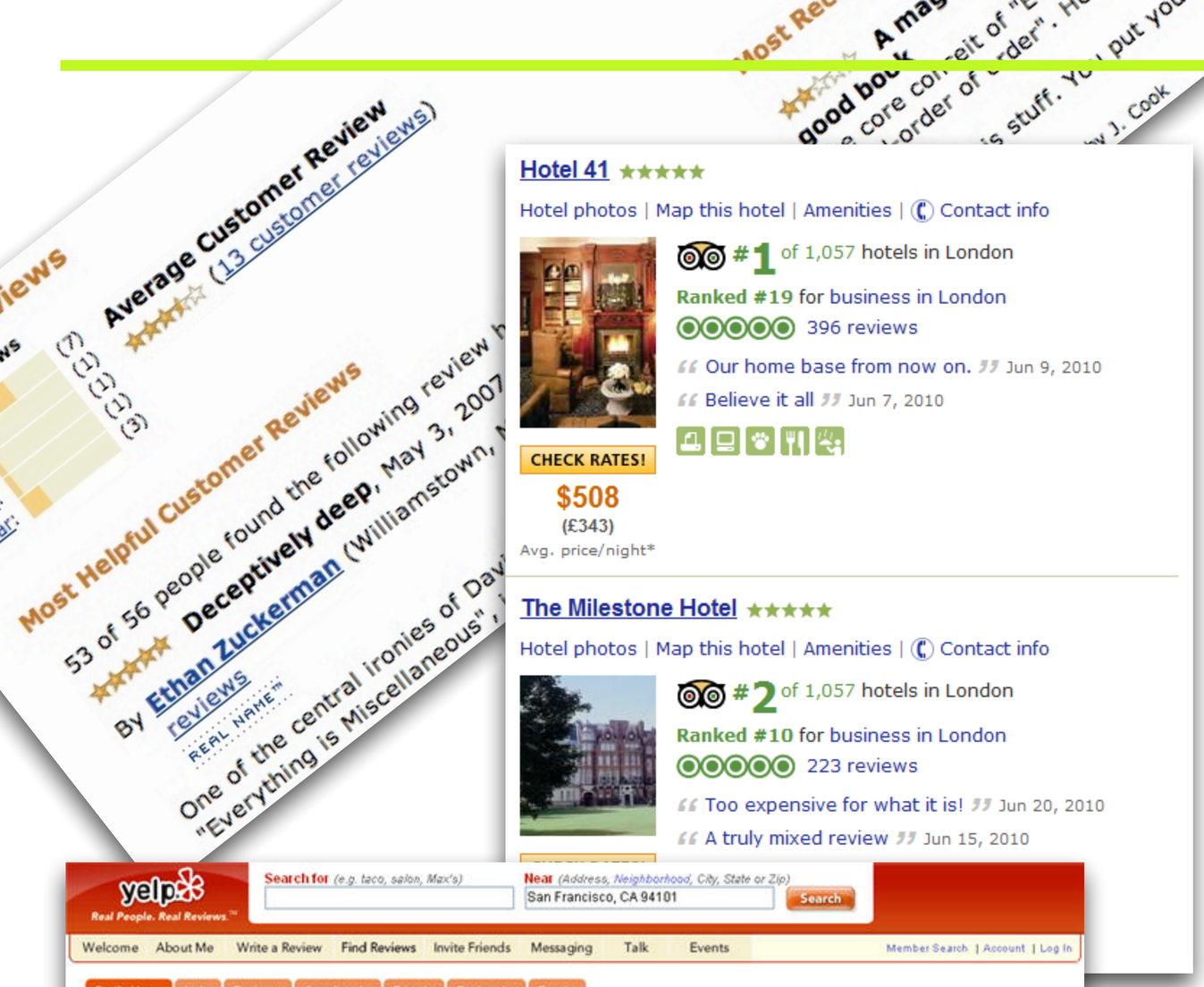
Schick and Schutze (2020), Exploiting cloze questions for Few Shot Text Classification and Natural Language Inference.

Solving NLP tasks with natural language inference (NLI)

Leveraging NLI models as universal solvers for diverse range of NLP tasks



Wang et al. (2021), Entailment as Few-Shot Learner.



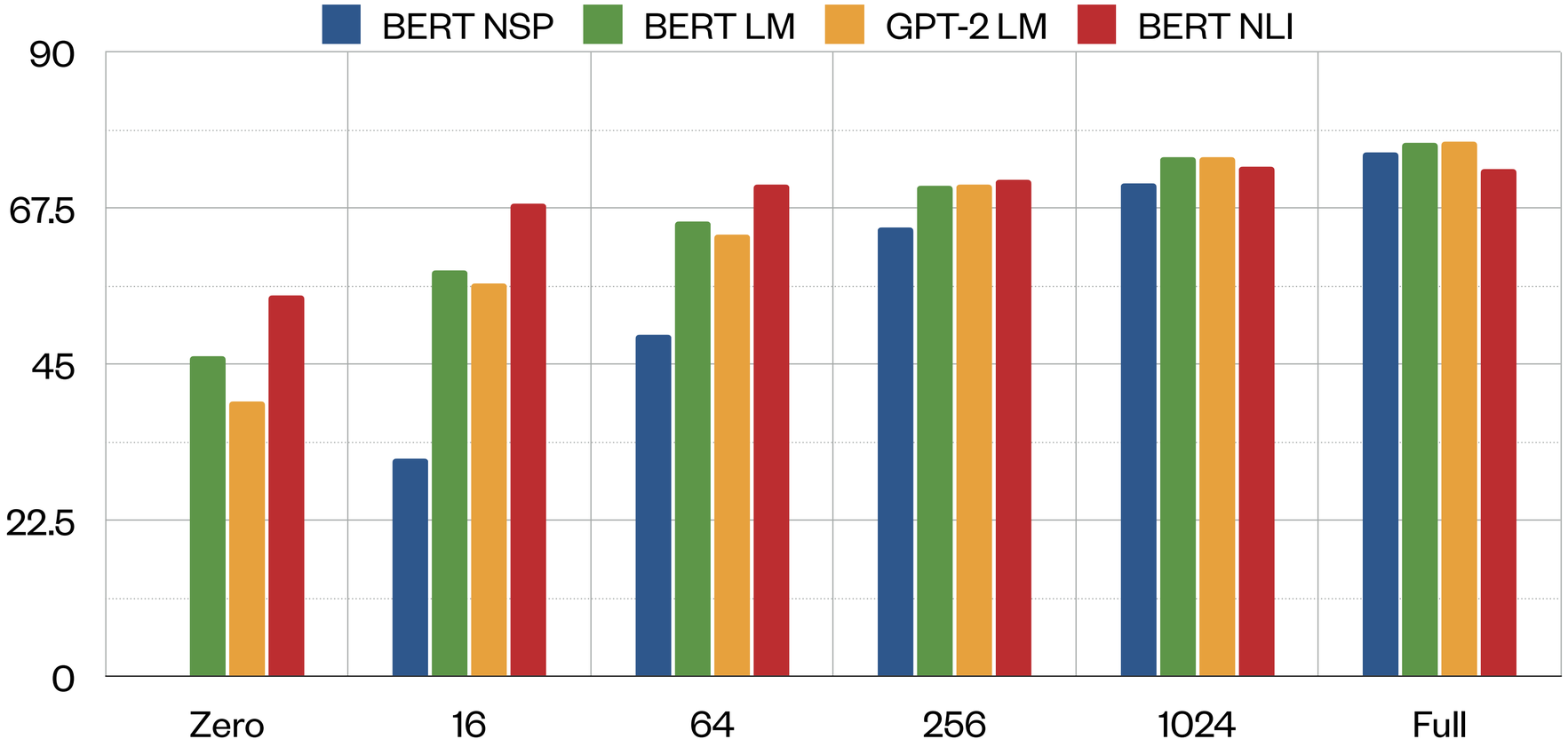
Why Prompts for ATSC?

- We could easily come up with prompt sentences that would naturally appear in review texts
  - Potential for LM/NLI's inherent abilities to perform ATSC
- Exploit large online shopping review corpora further
  - Usage of those corpora was limited to standard pretraining-finetuning scheme
  - Knowledge obtained from corpora could be better utilized when combined with prompts

Experiments

Prompts constantly outperform the no-prompt baselines.

- **Our zero-shot models perform better** than the baselines that did see some labeled examples.
- Larger performance gains achieved in **few-shot cases**
- **BERT NLI** does particularly well with lower number of examples
  - No in-domain pretraining done for NLI

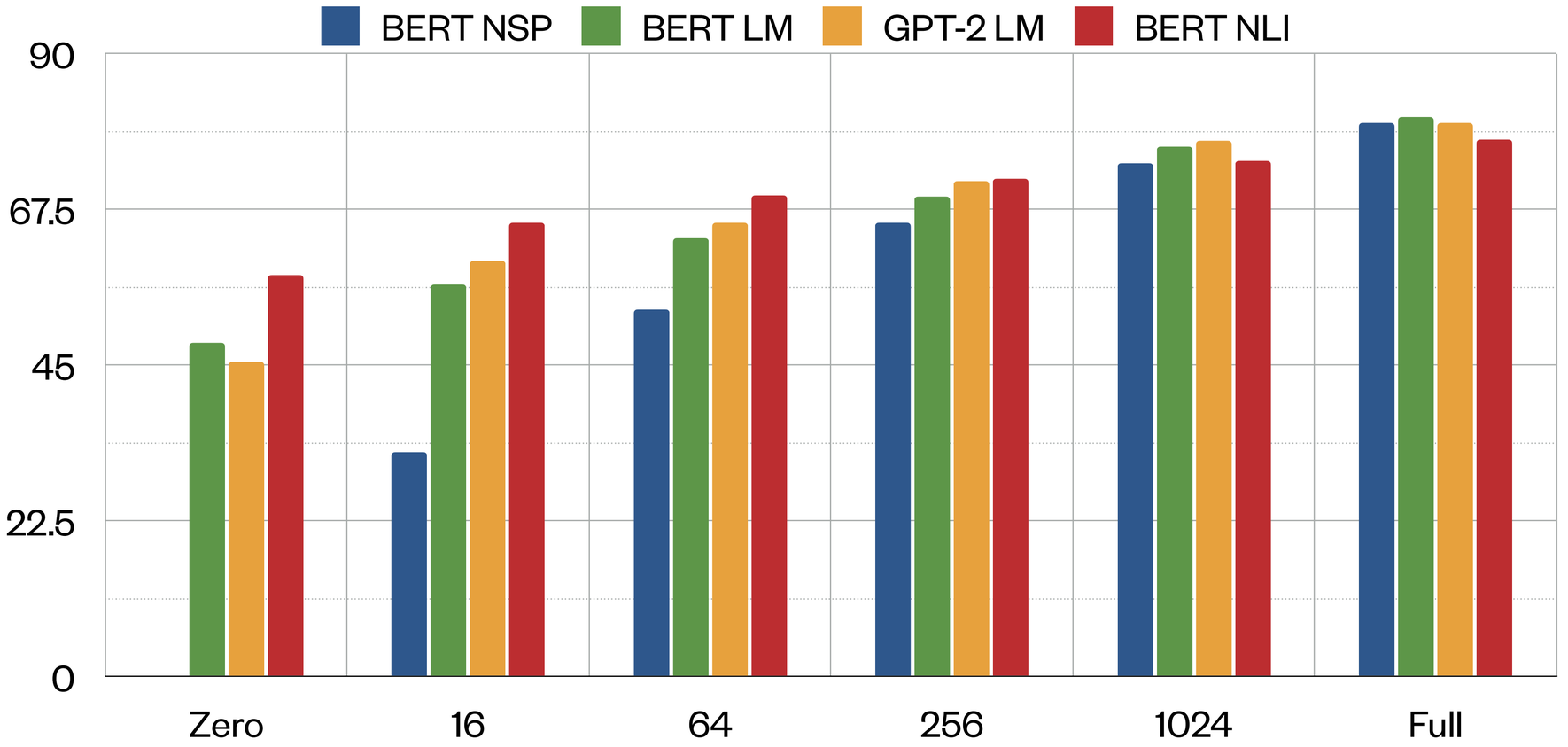


Macro F1 scores on SemEval 2014 ATSC Laptops

All scores are averaged over the 3 prompts and 5 random seeds.

Prompts constantly outperform the no-prompt baselines.

- **Our zero-shot models perform better** than the baselines that did see some labeled examples.
- Larger performance gains achieved in **few-shot cases**
- **BERT NLI** does particularly well with lower number of examples
  - No in-domain pretraining done for NLI

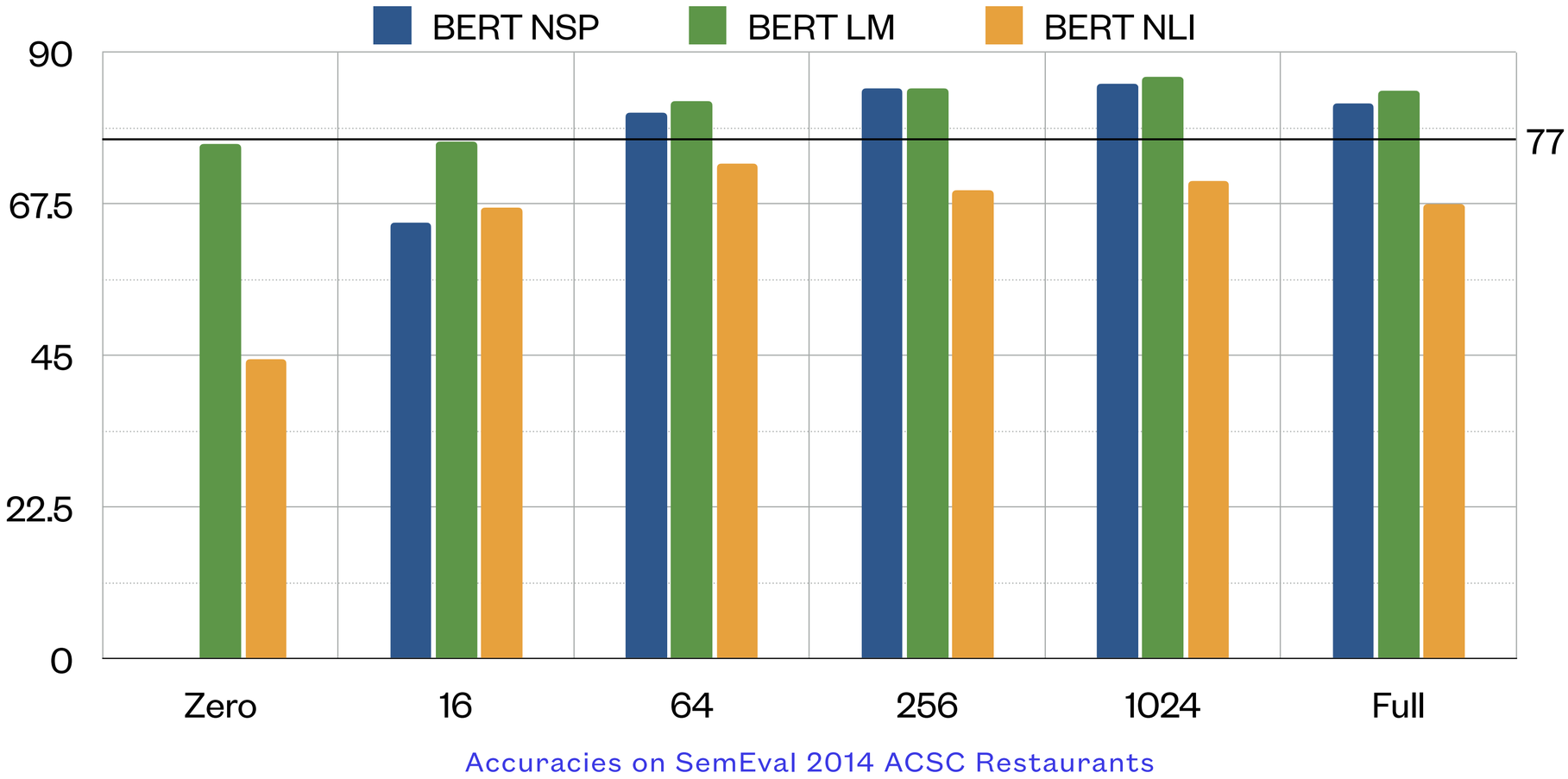


Macro F1 scores on SemEval 2014 ATSC Restaurants

All scores are averaged over the 3 prompts and 5 random seeds.

Prompts can better recognize implicit aspects.

- Test the models trained on ATSC on **ACSC without no extra training**
- Acquired some ability to recognize **aspects that are implied or worded differently** from the query aspect term.
- BERT NLI performs rather poorly, as it probably cannot recognize related domain-specific words

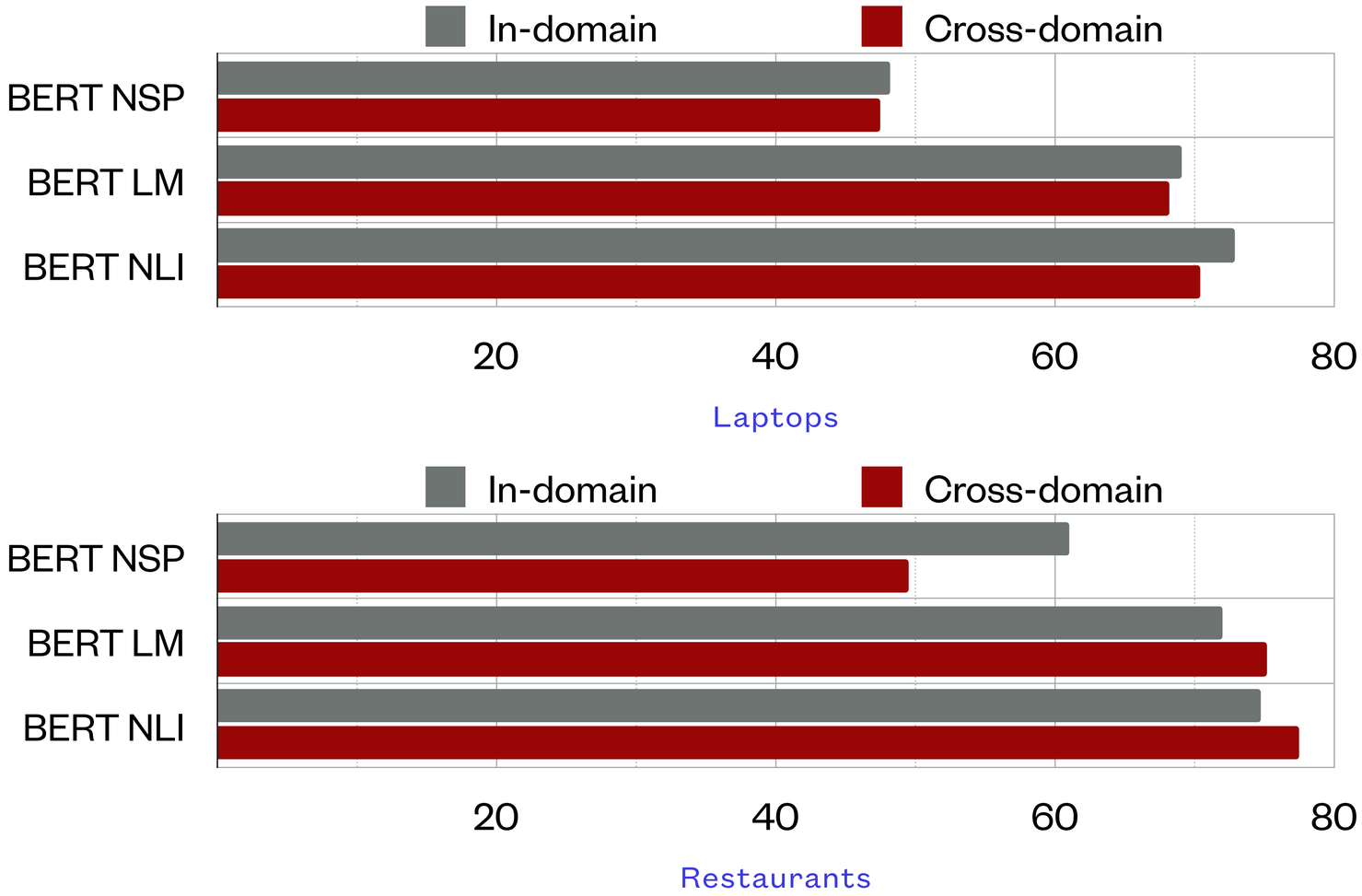


Accuracies on SemEval 2014 ACSC Restaurants

All scores are averaged over the 3 prompts and 5 random seeds.

Prompts can utilize cross-domain data effectively.

- With 16 cross-domain examples, BERT LM and NLI performs better than BERT NSP that were trained on *in-domain* examples.
  - Potential adaptability to **arbitrary domains** under low-resource settings



All scores are averaged over the 3 prompts and 5 random seeds.

Similar performance across different prompt choices

Practically reasonable choices of prompts could achieve good ATSC performance.

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

Micro F1 scores for target classes

- Observed better Micro F1 scores over the baseline in all classes
- BERT NLI is particularly better than LMs at discerning neutral and negative examples.

	Positive	Negative	Neutral
BERT NSP	63.65	22.76	7.64
BERT LM	83.51	60.83	31.46
GPT-2 LM	82.85	66.83	20.47
BERT NLI	**83.6**	**69.60**	**50.98**

Micro F1 scores for 16 examples from laptops domain

	Positive	Negative	Neutral
BERT NSP	75.50	14.47	7.40
BERT LM	86.34	57.65	25.95
GPT-2 LM	**87.95**	65.68	26.54
BERT NLI	80.86	**66.3**	**51.55**

Micro F1 scores for 16 examples from restaurants domain

Conclusion

Natural language prompts for ATSC

Make ATSC possible in **zero-shot** settings:
Conversion to NLI achieves considerable performance

Achieve superior **few-shot** learning performance:
Up to 33.14 MF1 improvements over the baselines

Possess abilities to handle **implicit** aspects:
Our models reach ~77% accuracy on ACSC with 16 ATSC examples

Going further

- Adapt our prompts to **jointly perform** aspect term extraction and sentiment classification (Luo et al. (2020))
- Explore potential ways of **combining ATSC, LM, and NLI** tasks into an unified task
  - In order to take the full advantage of both our unlabeled text and NLI pretraining
- Determine whether there are any strong **linguistic patterns** among the prompt models' predictions
  - Detailed insights into the potential behaviors of our prompt-based models

Luo et al. (2020), GRACE: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis.

Questions or feedback?

Please come to our poster session!

or send us an email to bseoh@cs.umass.edu