Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence

Chi Sun, Luyao Huang, Xipeng Qiu

Lorenzo Agnolucci

Supervisor: Prof. Paolo Frasconi

Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Firenze

Index

Introduction

Introduction

- In [Sun et al., 2019] (T)ABSA is converted to a sentence-pair classification task, such as question answering (QA) and natural language inference (NLI), by constructing an auxiliary sentence
- A pre-trained BERT model is fine-tuned to obtain new state-of-the-art results on SentiHood and SemEval-2014 Task 4 datasets
- Furthermore, a comparative experiment which relies on a single sentence classification approach is conducted to verify that the improvement is not only due to BERT but also to the proposed method

Task description

Sentiment Analysis (SA)

- Important task in natural language processing that allows to evaluate customer satisfaction on products and services from a given text
- The underlying assumption is that the entire text has an overall polarity
- Does not address different polarities associated to different aspects of a given entity

Fitting example:

"This book is so fun to read."

Unfitting example:

"This book is a hardcover version, but the price is a bit high."

Task description

Aspect-Based Sentiment Analysis (ABSA)

- Identifies fine-grained polarity towards a specific aspect
- Allows users to evaluate aggregated sentiments for each aspect of a given entity and gain a more granular understanding of its quality
- Can not handle sentences that refer to more than one target

Fitting example:

"The design of the space is good but the service is horrid!"

Unfitting example:

"The design of the space is good in Boqueria but the service is horrid, on the other hand, the staff in Gremio are very friendly and the food is always delicious."

Task description

Targeted Aspect-Based Sentiment Analysis (TABSA)

- Introduced by [Saeidi et al., 2016]
- Identifies fine-grained opinion polarity towards a specific aspect associated with a given target
- Two steps:
  1. Determine the aspects associated with each target
  2. Resolve the polarity of aspects to a given target

Example:

"location-1 is very safe and location-2 is too far"

BERT

- Bidirectional Encoder Representations from Transformers (BERT) is a language representation model introduced by [Devlin et al., 2018]
- BERT's model architecture is a multi-layer bidirectional Transformer encoder and is unified across different tasks
- BERT input representation unambiguously represents a single sentence or a pair of sentences in one token sequence. For a given token, its input representation is constructed by summing the corresponding token, segment and position embeddings
- BERT is pre-trained on unlabeled data on two unsupervised tasks: Masked Language Modeling and Next Sentence Prediction
- Fine-tuning BERT is straightforward: the model is initialized with the pre-trained parameters and then all the parameters are fine-tuned for a given task

Methodology

Auxiliary sentence construction

Four methods are considered to convert (T)ABSA into a sentence pair classification task:

QA-M

- The generated sentences are questions with the same format
- For example, for the target-aspect pair *(LOCATION1, safety)* the sentence is *"what do you think of the safety of location-1?"*

NLI-M

- The generated sentences are just pseudo-sentences with a much simpler form
- For example, for the target-aspect pair *(LOCATION1, safety)* the sentence is *"location-1 - safety"*

Auxiliary sentence construction (cont.)

QA-B

- TABSA is temporarily converted into a binary classification problem with $y \in \{yes, no\}$ to obtain the probability distribution
- Each target-aspect pair generates sequences such as "the polarity of the aspect safety of location-1 is positive", "the polarity of the aspect safety of location-1 is negative", "the polarity of the aspect safety of location-1 is none"
- The class of the sequence with the highest *yes* probability value is taken for the predicted category

NLI-B

- The only difference from QA-B is that the generated sentences are pseudo-sentences
- In this case, the auxiliary sentences are *"location-1 - safety - positive"*, *"location-1 - safety - negative"*, *"location-1 - safety - none"*

Fine-tuned models

BERT-single for (T)ABSA

- BERT for single sentence classification tasks
- With n_t and n_a number of target and aspect categories:
  - TABSA summarizes the results of $n_t \cdot n_a$ sentiment classifiers
  - ABSA combines the results of n_a sentiment classifiers

BERT-pair for (T)ABSA

- BERT for sentence-pair classification tasks
- One model for each auxiliary sentence construction method: BERT-pair-QA-M, BERT-pair-NLI-M, BERT-pair-QA-B, BERT-pair-NLI-B

Experiments

Datasets
SentiHood

- Introduced by [Saeidi et al., 2016] for TABSA
- Composed by 5215 sentences, 3862 with a single target and 1353 with multiple targets (already split in training, validation and test sets)
- Each sentence contains a list of target-aspect pairs $\{t, a\}$ with sentiment polarity y
- Given a sentence s and the target t in s the goal is to:
  1. detect the mention of an aspect a for the target t
  2. determine the sentiment polarity y for detected target-aspect pairs

Targets	*LOCATION1, LOCATION2*
Aspects	*general, price, safety, transit-location*
Polarities	*positive, negative, none*

Datasets

SemEval-2014 Task 4

- Introduced by [Pontiki et al., 2014] for ABSA
- Composed by 3841 sentences (already split in training and test sets)
- The only difference from the SentiHood dataset is that the target-aspect pairs $\{t, a\}$ become only aspects a
- This setting allows the joint evaluation of subtask 3 (Aspect Category Detection) and subtask 4 (Aspect Category Polarity)

Aspects	*food, price, service, ambience, anecdotes/miscellaneous*
Polarities	*positive, negative, conflict, neutral, none*

Implementation

- The implementation relies on the Hugging Face Transformers library and on PyTorch as backend

- The main pipeline for each model and dataset is:

1. Tokenize and encode the sentences of each example
2. Build a PyTorch *Dataset* with the encodings and the labels of all the examples
3. Load the chosen pre-trained BERT model as an instance of the built-in *BERTForSequenceClassification* class
4. Fine-tune the model using the built-in *Trainer* class with the specified hyperparameters
5. Use the fine-tuned model for inference on the test set and apply softmax to the outputted prediction scores
6. Evaluate model performance

Implementation
Hyperparameters

- Model: pre-trained uncased BERT-base
- Number of transformer blocks: 12
- Hidden layer size: 768
- Number of self-attention heads: 12
- Total number of parameters: 110M
- Dropout probability: 0.1
- Learning rate: 2e-5
- Warm-up proportion: 0.1
- Batch size: 24
- Number of epochs: 4

Metrics
SentiHood

- Aspect Detection:

  Given the example t of the dataset D, Y_t and $\widehat{Y}_t$ respectively ground-truth and predicted aspect categories:

  - *Strict accuracy* $= \frac{1}{|D|} \sum_{t \epsilon D} \sigma(Y_t = \widehat{Y}_t)$, with $\sigma(\cdot)$ indicator function
  - $Macro-F1 = \frac{2 \cdot Ma-P \cdot Ma-R}{Ma-P + Ma-R}$, with:
    - $Ma-P = \frac{1}{|D|} \sum_{t \epsilon D} \frac{|Y_t \cap \widehat{Y}_t|}{|\widehat{Y}_t|}$
    - $Ma-R = \frac{1}{|D|} \sum_{t \epsilon D} \frac{|Y_t \cap \widehat{Y}_t|}{|Y_t|}$
  - *AUC*: Uses prediction scores, with macro average and binary labels

- Sentiment classification:
  - In sentiment classification the scores of "None" are ignored
  - *Accuracy*: Number of correct polarities divided by the total number of polarities
  - *AUC*: Uses prediction scores, with macro average and binary labels

Results
SentiHood

Model	Aspect Detection			Sentiment Classification	
	Strict Acc.	Macro-F1	AUC	Accuracy	AUC
LR	-	39.3	92.4	87.5	90.5
LSTM-Final	-	68.9	89.8	82.0	85.4
LSTM-Loc	-	69.3	89.7	81.9	83.9
LSTM+TA+SA	66.4	76.7	-	86.8	-
SenticLSTM	67.4	78.2	-	89.3	-
Dmu-Entnet	73.5	78.5	94.4	91.0	94.8
BERT-single†	73.7	81.0	96.4	85.5	84.2
BERT-pair-QA-M†	79.4	86.4	97.0	**93.6**	96.4
BERT-pair-NLI-M†	78.3	87.0	**97.5**	92.1	96.5
BERT-pair-QA-B†	79.2	**87.9**	97.1	93.3	**97.0**
BERT-pair-NLI-B†	**79.8**	87.5	96.6	92.8	96.9
BERT-single	73.2	80.9	95.6	86.0	88.4
BERT-pair-QA-M	79.9	86.8	**97.3**	93.6	96.8
BERT-pair-NLI-M	78.2	86.4	96.8	93.0	96.2
BERT-pair-QA-B	**80.4**	**87.7**	96.9	**94.4**	**97.1**
BERT-pair-NLI-B	80.0	87.4	96.9	92.4	95.3

"-" means not reported, "†" means original result

Metrics

SemEval-2014 Task 4

- Aspect Category Detection:
  - $Micro-F1 = \frac{2 \cdot P \cdot R}{P + R}$
  - $Precision\ P = \frac{|S \cap G|}{|S|}$
  - $Recall\ R = \frac{|S \cap G|}{|G|}$

    with S set of predicted aspect categories and G set of ground-truth aspect categories
- Aspect Category Polarity:
  - 4-way: accuracy excluding "None" scores
  - 3-way: accuracy excluding "None" and "Conflict" scores
  - Binary: accuracy considering only "Positive" and "Negative" scores

Results

SemEval-2014 Task 4

Model	Aspect Category Detection			Aspect Category Polarity		
	Precision	Recall	Micro-F1	4-way	3-way	Binary
XRCE	83.23	81.37	82.29	78.1	-	-
NRC-Canada	91.04	86.24	88.58	82.9	-	-
LSTM	-	-	-	-	82.0	88.3
ATAE-LSTM	-	-	-	-	84.0	89.9
BERT-single†	92.78	89.07	90.89	83.7	86.9	93.3
BERT-pair-QA-M†	92.87	90.24	91.54	85.2	89.3	95.4
BERT-pair-NLI-M†	93.15	90.24	91.67	85.1	88.7	94.4
BERT-pair-QA-B†	93.04	89.95	91.47	**85.9**	**89.9**	**95.6**
BERT-pair-NLI-B†	93.57	90.83	**92.18**	84.6	88.7	95.1
BERT-single	93.42	88.59	90.94	82.2	86.3	93.2
BERT-pair-QA-M	93.43	90.24	**91.81**	85.2	89.2	94.8
BERT-pair-NLI-M	93.33	90.15	91.71	84.4	88.7	94.5
BERT-pair-QA-B	92.34	90.54	91.43	**86.1**	**89.5**	**95.2**
BERT-pair-NLI-B	94.12	89.07	91.53	85.9	89.4	94.1

"-" means not reported, "†" means original result

Observations

The improvement of the BERT-pair models w.r.t. the baselines is mainly due to two reasons:

1. Constructing an auxiliary sentence is equivalent to expanding the corpus. A sentence s_i in the original dataset will be expanded into $(s_i, t_1, a_1), \ldots, (s_i, t_1, a_{n_a}), \ldots, (s_i, t_{n_t}, a_{n_a})$ in the sentence-pair classification task

2. The improvement of BERT on the QA and NLI tasks ([Devlin et al., 2018]) shows that BERT has an advantage in dealing with sentence-pair classification tasks. This is the reason why directly fine-tuning BERT on (T)ABSA does not achieve a remarkable performance growth.

Conclusions

Conclusion

- In this paper (T)ABSA was converted to a sentence-pair classification task by constructing an auxiliary sentence
- Fine-tuning a pre-trained BERT model on this sentence-pair classification task allowed to achieve new state-of-the-art results
- A comparison with the results obtained by a model fine-tuned on single sentence classification showed how the proposed approach is the main reason behind the performance growth
- As a future work, the conversion method could be applied to other similar tasks

References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018).
BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR, abs/1810.04805.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (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.

Saeidi, M., Bouchard, G., Liakata, M., and Riedel, S. (2016).
Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods.
CoRR, abs/1610.03771.

Sun, C., Huang, L., and Qiu, X. (2019).
Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence.
CoRR, abs/1903.09588.

Appendix

Dataset composition: SentiHood

Train	Validation	Test	Total
2977	747	1491	5215

Sentences of the datasets

Aspect	Positive	Negative	Total
General	1610	446	2056
Price	356	514	870
Transit-location	609	156	765
Safety	267	328	595
Total	2842	1444	4286

Aspects distribution per sentiment class

Appendix

Dataset composition: SemEval-2014 Task 4

Train	Test	Total
3041	800	3841

Sentences of the datasets

Aspect	Positive	Negative	Conflict	Neutral	Total
Food	1169	278	82	121	1650
Price	230	143	20	11	404
Service	425	281	40	23	769
Ambience	339	119	60	31	549
Anecdotes/Misc.	673	240	45	408	1366
Total	2836	1061	247	594	4738

Aspects distribution per sentiment class