SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, Feng W

ACL 2020
TASK

- **Sentiment Analysis**: Identification of sentiment and opinion contained in the input texts that are often user-generated comments

  I like the quality of the food in this restaurant.

- **Sub-tasks**:
  - Sentence-level Sentiment Analysis: Predict sentiment polarity for a given sentence
  - Aspect-based Sentiment Analysis: Predict sentiment polarity with respect to a specific aspect (term/category)
  - Opinion Word Extraction: Identifying the words conveying opinion
  - Aspect Target Detection: Recognizing the words that are target of an opinion phrase

- **Feature-based or task specific architectures** were extensively used in prior works

- **Using pre-trained language models**, BERT, are the basis for SOTA models
Contributions

● Define three different sentiment-based objective in pre-training of RoBERTa
  ○ Sentiment-Pair prediction: Predict masked aspect term and sentiment word
  ○ Sentiment Word Prediction: Predict the gold masked sentiment word
  ○ Word Polarity Prediction: Predict the sentiment polarity of the masked words

● None of the prior works have considered sentiment-based pretraining

● Evaluate the pre-trained model on:
  ○ Sentence-level sentiment prediction
  ○ Aspect-based sentiment analysis
  ○ Opinion word extraction

● Provide analysis on contribution of each task

● Results show the pre-trained model improve performance and results in more reasonable attention weights for sentiment analysis in RoBERTa
Model Overview
Background

- BERT is a pre-trained language model which employs encoder of Transformer
- Objective:
  - Masked Language Modeling: Random words in inputs are replaced by [MASK] and their corresponding embeddings are used to predict the original word
  - Next Sentence Prediction: Representation of the sentence is used to predict if two pair of sentences are consecutive or not
- After pre-training, a task-specific head is added and the model is fine-tuned on labeled data
- RoBERTa:
  - Employs more robust optimization compared to BERT
  - Remove next sentence prediction task in pre-training
  - Achieves better results
Sentiment Knowledge Enhanced Pre-training

- Given an unlabeled task-specific corpus $X$, the masked version $X'$ is created such that sentiment knowledge in $X'$ is masked
  - Sentiment Words
  - Sentiment Polarity of words
  - Aspect-Sentiment Pairs
- The transformer-based language model (i.e., RoBERTa) is pre-trained on the masked dataset $(X, X')$
- Pre-trained model is further fine-tuned on the labeled datasets for tasks:
  - Sentence level Sentiment Analysis
  - Aspect-based Sentiment Analysis
  - Opinion Word Analysis
Constructing Sentiment Knowledge

● Compute the sentiment polarity of words in the vocabulary
  ○ A small seed of words are manually selected
  ○ The Pointwise Mutual Information (PMI) between all words and seed words are computed:
    \[ PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \]
  ○ The difference between PMI with positive and negative seed words is used to automatically identify positive and negative words in the vocabulary
    \[ WP(w) = \sum_{WP(s)=+} PMI(w, s) - \sum_{WP(s)=-} PMI(w, s) \]

● Find sentiment-aspect pairs:
  ○ Every noun in proximity of the sentiment words (i.e., no more than 3 word away) is selected as aspect term
Sentiment Masking

- **Sentiment Detection**: Recognize which words or pairs in the input sentences in the corpus convey sentiment
  - Sentiment Word Detection: If a word of the sentence is included in the list of positive or negative words constructed in the previous step, this word is selected as sentiment word
  - Aspect-sentiment pair: Similar to sentiment knowledge construction, if a noun appears in the proximity of a sentiment word it is selected as aspect term

- **Remove sentiment information from input text**:
  - At most 2 sentiment-pairs are randomly selected and replaced with [MASK]
  - 10% of sentiment words are replaced with [MASK]
  - If there are not enough sentiment word, common words are used

- **The constructed masked corpus, X’**, is used to pre-train RoBERTa
Pre-training Objectives

- Three tasks as objectives of pre-training:
  - Sentiment Word Prediction
    \[ \hat{y}_i = \text{softmax}(\tilde{x}_i \mathbf{W} + \mathbf{b}) \]
    \[ L_{sw} = - \sum_{i=1}^{i=n} m_i \times y_i \log \hat{y}_i \]
  - Word Polarity Prediction: Predict the polarity of masked sentiment word (binary prediction)
  - Sentiment-Pair Prediction: Simultaneously predict words of masked sentiment-pairs (multi-label classification)
    \[ \hat{y}_a = \text{sigmoid}(\tilde{x}_1 \mathbf{W}_{ap} + \mathbf{b}_{ap}) \]
    \[ L_{ap} = - \sum_{a=1}^{a=A} y_a \log \hat{y}_a \]
Tasks

- Three tasks for fine-tuning
  - Sentence-level Sentiment Analysis: The representation of [CLS] is used to predict the sentiment of the input text
  - Aspect-based sentiment analysis: The input is fed into the pre-trained model in the form of [CLS] sentence [SEP] aspect. The representation of [CLS] is used for prediction
  - Opinion Role Labeling: The holder and target of the opinion words are predicted. It is formulated as sequence labeling with CRF as the head.

```plaintext
Australia said [it]_H [feared]_{O_{neg}} [violence]_T
if voters thought the election had been stolen.
```
Datasets

- Five datasets are used for fine-tuning
- Amazon-2 is used for pre-training
- Models with different random seed and parameters are evaluated on development set and the one with medium performance is evaluated on test set
Main Results

- Proposed method improves performance of base and large versions of RoBERTa
- Results are SOTA on all except one dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Sentence-Level</th>
<th>Aspect-Level</th>
<th>Opinion Role</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SST-2</td>
<td>Amazon-2</td>
<td>Sem-L</td>
</tr>
<tr>
<td>Previous SOTA</td>
<td>97.1(^1)____</td>
<td>97.37(^2)</td>
<td>81.35(^3)</td>
</tr>
<tr>
<td>RoBERTa(_{base})</td>
<td>94.9</td>
<td>96.61</td>
<td>78.11</td>
</tr>
<tr>
<td>RoBERTa(_{base}) + SKEP</td>
<td>96.7</td>
<td>96.94</td>
<td>81.32</td>
</tr>
<tr>
<td>RoBERTa(_{large})</td>
<td>96.5</td>
<td>97.33</td>
<td>79.22</td>
</tr>
<tr>
<td>RoBERTa(_{large}) + SKEP</td>
<td>97.0</td>
<td>97.56</td>
<td>81.47</td>
</tr>
</tbody>
</table>
Ablation Study

- All three tasks are necessary for best performing pre-trained model
- Predicting sentiment-pair words independently (instead of multi-label classification) has lower performance on most datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>Sentence-Level</th>
<th></th>
<th>Aspect-Level</th>
<th></th>
<th>Opinion Role</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SST-2 dev</td>
<td>Amazon-2</td>
<td>Sem-L</td>
<td>Sem-R</td>
<td>MPQA-Holder</td>
</tr>
<tr>
<td>RoBERTa&lt;sub&gt;base&lt;/sub&gt;</td>
<td>95.21</td>
<td>96.61</td>
<td>78.11</td>
<td>84.93</td>
<td>81.89/77.34</td>
<td>80.23/72.19</td>
</tr>
<tr>
<td>+ Random Token</td>
<td>95.57</td>
<td>96.73</td>
<td>78.89</td>
<td>85.77</td>
<td>82.71/77.71</td>
<td>80.86/73.01</td>
</tr>
<tr>
<td>+ SW</td>
<td>96.38</td>
<td>96.82</td>
<td>80.13</td>
<td>86.92</td>
<td>82.95/77.63</td>
<td>81.18/73.15</td>
</tr>
<tr>
<td>+ SW + WP</td>
<td>96.51</td>
<td>96.87</td>
<td>80.32</td>
<td>87.25</td>
<td>82.97/77.82</td>
<td>81.09/73.24</td>
</tr>
<tr>
<td>+ SW + WP + AP</td>
<td>96.87</td>
<td>96.94</td>
<td>81.32</td>
<td>87.92</td>
<td>84.25/79.03</td>
<td>82.77/74.82</td>
</tr>
<tr>
<td>+ SW + WP + AP-I</td>
<td>96.89</td>
<td>96.93</td>
<td>81.19</td>
<td>87.71</td>
<td>84.01/78.36</td>
<td>82.69/74.36</td>
</tr>
</tbody>
</table>
Analysis

- Predicting words of a pair using [CLS] or corresponding word representations

<table>
<thead>
<tr>
<th>Model</th>
<th>SST-2 dev</th>
<th>Sem-L</th>
<th>Sem-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent-Vector</td>
<td>96.87</td>
<td>81.32</td>
<td>87.92</td>
</tr>
<tr>
<td>Pair-Vector</td>
<td>96.91</td>
<td>81.38</td>
<td>87.95</td>
</tr>
</tbody>
</table>

- Attention scores for computing [CLS] representation in the final layer

<table>
<thead>
<tr>
<th>From</th>
<th>Model</th>
<th>Sentence Samples</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>RoBERTa</td>
<td>altogether, this is successful as a film, while at the same time being a most touching reconsideration of the familiar masterpiece.</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>SKEP</td>
<td>altogether, this is successful as a film, while at the same time being a most touching reconsideration of the familiar masterpiece.</td>
<td>positive</td>
</tr>
<tr>
<td>Sem-L</td>
<td>RoBERTa</td>
<td>I got this at an amazing price from Amazon and it arrived just in time.</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>SKEP</td>
<td>I got this at an amazing price from Amazon and it arrived just in time.</td>
<td>positive</td>
</tr>
</tbody>
</table>
Conclusion

- A sentiment-based pre-training approach for RoBERTa is proposed
- Sentiment knowledge construction is automatic
- Task specific corpus is employed for fine-tuning
- Pre-trained model is evaluated on three different tasks
- Experiments show all tasks of pre-training are helpful to improve performance
- SOTA results is achieved on various datasets
- Attention scores of RoBERTa are more meaningful using this approach