Main Tasks

Task 1: Sentence classification
For example: error detection, sentiment analysis, hedge detection.

It was so long time to wait in the theatre.
Therefore, houses will be built on high supports.

Task 2: Token labeling
Many of these tasks can also be performed on the token level.
I like to playing the guitar and sing louder.

Idea: Join these two objectives together into one model, such that they start helping each other.

Supervised Self-attention

- Bi-LSTM predicts label confidence scores \( a_i \) for each input token.
- The same scores are used as self-attention weights to predict the sentence label \( y \).

\[
\begin{align*}
\text{Sentiment} & = \sum_{i=1}^{N} a_i h_i \\
\text{Task 2} & = \sum_{i=1}^{N} \left( a_i - \hat{a}_i \right)^2 \\
\text{Jointly} & \quad \text{learn token labeling} \quad \text{and better word representations.}
\end{align*}
\]

- Sigmoid-activated attention weights allow the system to predict multiple positive values in a sentence:
\[
a_i = \sigma(W_{a} e_i + b_a) \quad a_i = \frac{a_i}{\sum_{k=1}^{N} a_k} \quad s = \sum_{i=1}^{N} a_i h_i
\]

- The model is jointly optimized for both sentence classification and token labeling:
\[
L_{\text{sent}} = \sum (y^{(i)} - \hat{y}^{(i)})^2 \quad L_{\text{task}} = \sum_{i} \left( a_i^{(i)} - \hat{a}_i^{(i)} \right)^2
\]

- Directly training the model to focus more on the words that human annotators found to be important.

Language Modeling Objectives

- In addition to the main objectives, predicting the previous and the next word in the sequence at each step, based on Rei (2017).
- Extending the method also to characters, optimizing a character-level Bi-LSTM for composing word representations.

\[
\begin{align*}
\text{Language} & = \sum_{i=1}^{N} f_i h_i \\
\text{Modeling} & = \sum_{i} \left( f_i - \hat{f}_i \right)^2
\end{align*}
\]

- Helps the model learn better word representations and better composition functions, thereby improving performance on both classification tasks.

Sentence Classification Experiments

- Evaluating on the tasks of hedge detection (CoNLL 2010), error detection (FCE) and sentiment analysis (Stanford Sentiment Treebank).

- The self-attention architecture (BiLSTM-ATTN) by itself has a slight advantage over the basic sentence classifier (BiLSTM-LAST).

- Including the token-level and language modeling objectives (BiLSTM-JOINT) considerably improves performance on all tasks.

Token Labeling Experiments

- The model is able to learn token labeling from the sentence-level objective, without having the whole dataset token-annotated.

- Achieves good performance even when no token-level annotation is available: 76.5% \( F_1 \) on CoNLL 2010.

- Maintains advantage also when the whole dataset is annotated.

Conclusion

- We can jointly train the model for sentence classification and token labeling, improving performance on both tasks.

- The language modeling objectives help learn better language representations for both levels of classification.

- The resulting model is a robust text classifier that is able to point to individual words in the sentence to justify its decisions.

Related Papers
