Jointly Learning to Label Sentences and Tokens

Marek Rei

Anders Søgaard
Task 1: Sentence Classification

Error Detection

It was so long time to wait in the theatre.
I like to playing the guitar and sing very louder.
This is a great opportunity to learn more about whales.
Therefore, houses will be built on high supports.

Sentiment Analysis

The whole experience exceeded our expectations.
Tom Hanks gave a fantastic performance as the lead.
Sundance fans always try to find the Next Great Thing.
The movie takes some time to come to the conclusion.
Task 2: Sequence Labeling

Error Detection

- - - - X - - - - - - X -

I like to **playing** the guitar and sing very **louder**.

Sentiment Analysis

- - - - - - X - - - - - -

Tom Hanks gave a **fantastic** performance as the lead.
Main Idea

01. Join together predictions on both sentences and tokens

02. Token-level predictions act as self-attention weights

03. Teaching the model where it should be focusing in the sentence
Model Architecture

Make token-level prediction scores also function as sentence-level attention weights.

\[ s = \sum_{i=1}^{N} a_i h_i \]
\[ L_{sent} = \sum_t (\hat{y}^{(t)} - y^{(t)})^2 \]
\[ L_{tok} = \sum_t \sum_i (\hat{a}_i^{(t)} - a_i^{(t)})^2 \]
Soft Attention Weights

Based on sigmoid + normalisation:

\[
\tilde{a}_i = \frac{1}{1 + \exp(-\tilde{e}_i)}
\]

Token-level prediction

\[
a_i = \frac{\tilde{a}_i}{\sum_{k=1}^{N} \tilde{a}_k}
\]

Self-attention weight

We can constrain the attention values based on the sentence-level label.
Language Modeling Objectives

1. Jointly training the network as a language model.

   Predicting the previous and the next word in the sequence.

2. Same principle extended to characters.

   Predicting the middle word based on characters of the surrounding words.
Evaluation

**CoNLL 2010** (Farkas et al., 2010)

Detecting speculative (hedged) language.
Shared task dataset, containing sentences from biomedical papers.

**FCE** (Yannakoudakis et al., 2011)

Detecting grammatically incorrect phrases and sentences.
Error-annotated essays written by language learners.

**Stanford Sentiment Treebank** (Socher et al., 2013)

Detecting sentiment in movie reviews.
Split into positive and negative sentiment detection.
Results: Sentence Classification

Supervision on the token level explicitly teaches the model where to focus for sentence classification.
Results: Sequence Labeling

Supervision on the sentence level regularizes the sequence labeler and encourages it to predict jointly consistent labels.

![Graph showing performance on CoNLL10 and FCE datasets with BiLSTM-ATTN and BiLSTM-JOINT models.](image)
Conclusion

01. Token-level labels can be used to supervise the attention module for sentence-level composition

02. Sentence-level labels can be used to regularize the token-level predictions

03. Language modeling objectives on tokens and characters help the model learn better composition functions

04. The result is a robust sentence classifier that is able to point to individual tokens to explain its decisions
Thank you!
Any questions?