Evaluating Natural Language Models on Technical Query Tagging

Atindra Jha
(atj10@stanford.edu)

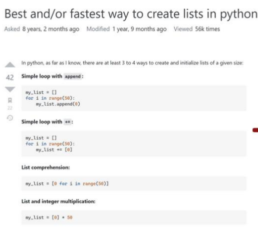
John Dalloul
(jdalloul@stanford.edu)

Kantapong Kotchum
(kkotchum@stanford.edu)

Overview

Given Stack Overflow Exchange questions and titles, predict the most likely question tags

Question Title + Body



Tags

python list

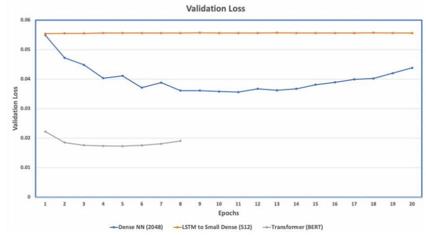
Data

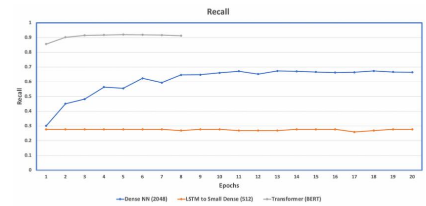
Our question/tag dataset consists of 300k questions, each of which is assigned a list of one or more tags. Additionally, we performed the following preprocessing steps:

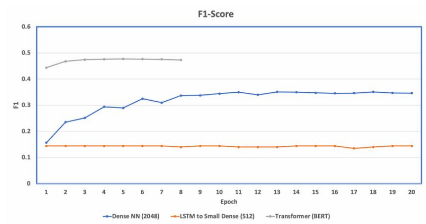
- removal of html tags & accent marks
- expansion of grammatical abbreviations
- conversion to lowercase

Our data was divided into train/val/test splits in the ratio of 80/10/10.

Charts







Results

Model	Validation			Test		
	Loss	Recall	F1 Score	Loss	Recall	F1 Score
Dense NN (2048)	0.04017	67.32%	0.3509	0.0408	65.48%	0.3478
BiDirectional LSTM	0.05558	27.64%	0.1441	0.0578	24.45%	0.1308
Transformer (BERT)	0.01726	92.01%	0.4772	0.0179	90.32%	0.4697

The trajectories of Binary Cross-Entropy Loss, Recall, and F1 score across training epochs are shown in the graphs to the left. In the table above, the best validation metrics are listed alongside statistics calculated on the held-out test set, if assessed.

Analysis

Dense Neural Network

- Input is fully represented (335 max words x 300 dimensions)
- Attains high recall and F1 scores, suggesting that the network excels at outputting tags that are accurate for the question

Bidirectional LSTM

- Sequential data input is ill-fitted for our task
- The final, dense layers receive a lower-dimensional representation of the input (256 hidden-unit dimensions)

Bidirectional Encoder Representations from Transformers (BERT)

- More robust word embeddings and attention can better grasp inter-word relationships
- Highest recall and F1 scores, indicating the best task performance