Stance Classification with LSTM-CNN Model

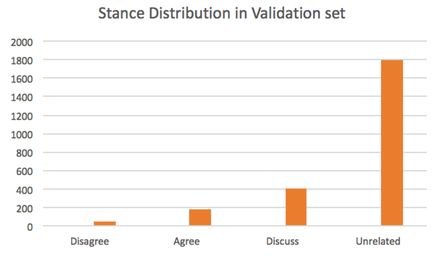
Ye Wang
University of Waterloo

Dataset

Training Dataset
Unique 66,677 headlines and 2587 article bodies, headlines and article bodies are either from the same news article or from two different articles.

Development Dataset
The final validation dataset used for tunning model has total 2438 headline-body pairs.

Goal: stance Classification
Labels:
66,677 headline-body pairs are labeled with four categories.
- "Agree"
- "Disagree"
- "Discuss"
- "Unrelated"



Data Cleaning & Features Extraction

Data Cleaning
1. Letters to lowercase
2. Filter punctuation
3. Remove stop words
4. Remove non-alphabetic characters
5. Stemming
6. Tokenization

Features Extraction
- Overlap Features
- Refuting Features
- Polarity Features

Text Pairs Representation

Training data representations are gained by averaging word vectors in each headline-bodies pair and then concatenating with the additional features into a single vector after the preprocessing step. This set of features have shown to be the best comparing to using word2vec features on their own or any of the reduced combinations of these features.

Experimental Setup & Evaluation

Hyper-parameter Selection

- The number of LSTM layers for headline is one or two
- The number of CNN layers for bodies is six
- The mini-batch size is either 32, 64, or 128
- The dropout probability is 0.2
- The number of epochs is selected from {50, 100, 200, 300}

Evaluation
All experiments were processed through minimizing the categorical crossentropy loss on the training set and monitoring the performance on the development set.

- Categorical accuracy (build-in function in keras)
- Precision score
- F1 score
- Recall score
- Macro F1

Use word2Vec word embedding in development dataset:

Methods

1. Bag-of-Words Model
2. Long Short Term Memory Network (LSTM)
3. Convolutional Neural Network Model (CNN)
4. LSTM-CNN with hand-crafted Features Model

Methods	Word2Vec	GloVe
Bag-of-Words	N/A	N/A
CNN	Y	Y
LSTM	Y	Y
LSTM-CNN Model	Y	Y

Original Data Results

Model	Train Loss	Stance F1 Score				Categorical Accuracy	Average Precision	Average Recall	Macro F 1
		Disagrees	Agrees	Discuss	Unrelated				
BoW	0.267	0.396	0.610	0.837	0.946	0.863	0.562	0.746	0.700
LSTM	0.129	0.490	0.734	0.821	0.958	0.896	0.727	0.798	0.751
CNN	0.141	0.427	0.759	0.794	0.926	0.870	0.690	0.777	0.730
LSTM-CNN	0.163	0.660	0.813	0.835	0.950	0.894	0.866	0.810	0.815

Oversampling Data Results

Model	Train Loss	Stance F1 Score				Categorical Accuracy	Average Precision	Average Recall	Macro F 1
		Disagrees	Agrees	Discuss	Unrelated				
BoW	0.179	0.479	0.627	0.815	0.954	0.881	0.720	0.790	0.719
LSTM	0.089	0.698	0.714	0.826	0.967	0.921	0.822	0.849	0.801
CNN	0.091	0.620	0.705	0.812	0.931	0.905	0.791	0.820	0.767
LSTM-CNN	0.063	0.758	0.827	0.920	0.972	0.950	0.866	0.863	0.870

Conclusions

In this project, our LSTM-CNN model is well performed. Our work demonstrates the efficiency of hand-crafted features and a oversampling method for distinguishing the classes with lower training data examples. In fact, the imbalanced training data for stance classification is a big challenge for tunning parameters. We have tried many times in different number of training dataset for testing the performance with various parameter combinations which needs lots of time and energy.

In the future, we will try to use Tree of Parzen Estimators (TPE) algorithm to search the parameter space which can find the most appropriate parameters combination more efficiently.

Future Works

Data Preprocessing
To counter class imbalance problems, normal methods we could apply in data level is divided in to the following two parts:
1. Undersampling
2. Cluster-Based over sampling

Parameters Exploration
In terms of sensitivity analyses, we could retrain our models with more epochs and then tend to explore certainly the relationship between batch size, the number of datasets and the number of epochs.

Model Extensions
We could explore bidirectional LSTMs in conditional encoding with attention mechanism.

Acknowledgments
We would like to acknowledge and thank MSCI 641 Professor Olga Vechtomova and TA Hareesh Pallikara Bahuleyan for their mentorship on this project.