Movie Review Sentiment Analysis

Machine Learning for Language Processing Project 2018/2019

1 Project Description

The aim of this project is to implement a machine learning model for a sentiment analysis task using the Rotten Tomatoes movie review dataset. During this project you are asked to label phrases on a scale of five values :

0. negative
1. somewhat negative
2. neutral,
3. somewhat positive,
4. positive.

Obstacles like sentence negation, sarcasm, terseness, language ambiguity, and many others make this task very challenging.

2 Data Description

The dataset is a corpus of movie reviews originally collected by Pang and Lee [1].This dataset contain tab-separated files with phrases from the Rotten Tomatoes dataset. The data are splitted to **train/test** sets and the sentences are shuffled from their original order.

— Each Sentence has been parsed into many **phrases** by the Stanford parser.
— Each phrase has a PhraseId.
— Each sentence has a **SentenceId**.
— Phrases that are repeated (such as short/common words) are only included once in the data.

The training set contain 156000 examples and the test set represent 66300 phrases. In the following table you can find several phrases and their Sentiment score.

PhraseId	SentenceId	Phrase	Sentiment
8140	336	of inept filmmaking	1
8143	336	joyless , idiotic , annoying , heavy-handed ,	0
8146	336	joyless	0
8147	336	idiotic , annoying , heavy-handed	2

TABLE 1 – Example of phrases from the training data and their scores

Data are available at :

train : `http://perso.univ-lemans.fr/~fbouga/train.tsv`
test : `http://perso.univ-lemans.fr/~fbouga/test.tsv`

3 Evaluation

Systems are evaluated on classification accuracy (the percent of labels that are predicted correctly) for every parsed phrase.

4 Project Roadmap

1. Study and plot the training data
2. Split the data (train/dev)
3. Train and evaluate a vanilla deep recurrent neural network (RNN)
4. Use the Pytorch framework to train the RNN network.
5. Optimize the model and propose enhancement (Regularization, network init, new architecture)
6. Prepare the final defense