Sentiment Analysis

Taixing Bi

Sentiment Analysis in Finance Domain

- Financial data is a challenging use case for Bullish or Bearish Sentiment Analysis
- Sentiments and opinions can affect market dynamics.

Datasets

- Prediction Data: earnings call transcripts 2014-2018
- Size: 20M

Outline

- Learning cross domain word embedding

  . Word embedding

  . word semantics in cross domain

  . Algorithm

  . Process

- Sentiment model

  . Hierarchical attention networks

  . Neural network architecture

  . Training and metrics

Word Embeddings

- Word Embeddings(learning of distributed representations for natural language words) has received a significant amount of attention recently
- Such representations were shown to be able to capture syntactic and semantic level information associated with words
- Effective in tasks:

  . Sentiment analysis

  . Text similarity

  . Named entity recognition(NER)

Why not train word embedding using earning call directly?

- Earnings call is 20 M. Coverage(accuracy) is only round 50%.
- Load pretained word embedding model, GoogleNews or GLove.

Size of Training	base tri-gram coverage	4 skip tri-gram coverage
10 M words	44.36%	53.76%
27.5 M words	53.23%	62.59%
50 M words	60.16%	**69.04%**
100 M words	65.31%	74.18%
200 M words	**69.37%**	79.44%

Reference

[1] A Closer Look at Skip-gram Modelling

Word Semantics in Cross Domain

- Word in different domain “semantics could be different”

O**ption**

. Google dictionary:

Option is a thing that is or may be chosen

. Finance dictionary:

Option is a contract which gives the buyer (the owner or holder of the option) the right, but not the obligation, to buy or sell an underlying asset or instrument at a specified strike price on a specified date, depending on the form of the option

Algorithm

$$\mathcal{L}'_{\mathcal{D}_t} = \mathcal{L}_{\mathcal{D}_t} + \sum_{w \in \mathcal{D}_t \cap \mathcal{D}_s} \alpha_w \cdot ||\mathbf{w}_t - \mathbf{w}_s||^2$$

$$\alpha_w = \sigma(\lambda \cdot \phi(w))$$

$$\phi(w) = \frac{2 \cdot \mathcal{F}_{\mathcal{D}_s}(w) \cdot \mathcal{F}_{\mathcal{D}_t}(w)}{\mathcal{F}_{\mathcal{D}_s}(w) + \mathcal{F}_{\mathcal{D}_t}(w)}$$

$\mathcal{L}_{\mathcal{D}_t}$ is skip-gram

$\phi(w)$ is word similarity in based on frequency in target corpus (earnings call) and source corpus (GloVe)

Reference

[1] A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings

[2] A Closer Look at Skip-gram Modelling

[3] word embedding-my GitHub

Process

- step 1 Load pre-trained model based on GloVe
- step 2. Generate similarity $\phi(w)$ score based on common words in earnings call and GloVe
- step 3. Training Cross-Domain Word Embeddings algorithm

Implantation language C

Result: [my github]

Words not in dictionary when training

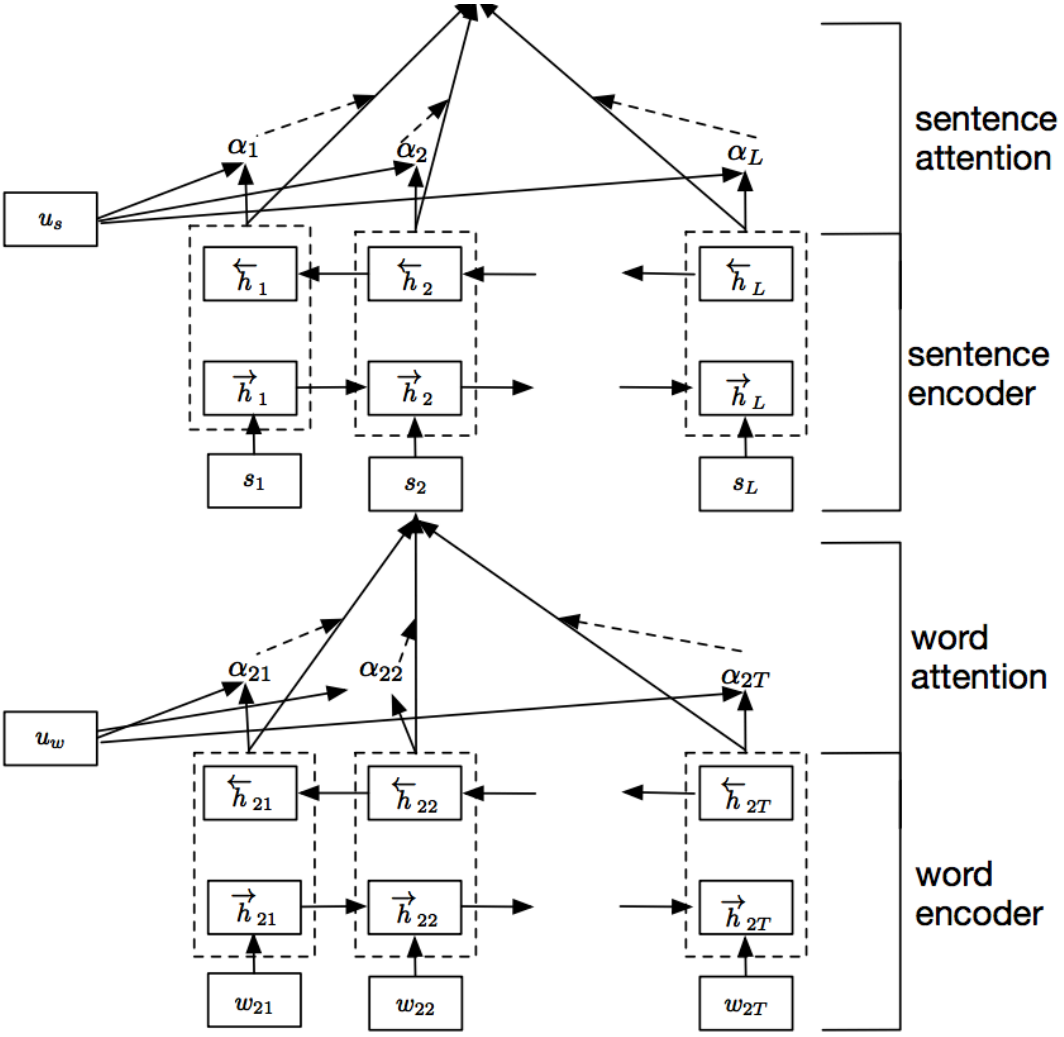
if words not in found in dictionary in GLoVe.

- Solution 1 randomly initialize weight (-0.25,0.25)
- Solution 2 n-grams char

  For instance: Antidisestablishmentarianism(not in GLoVe)

  n-grams char: Antidise (in GLoVe)

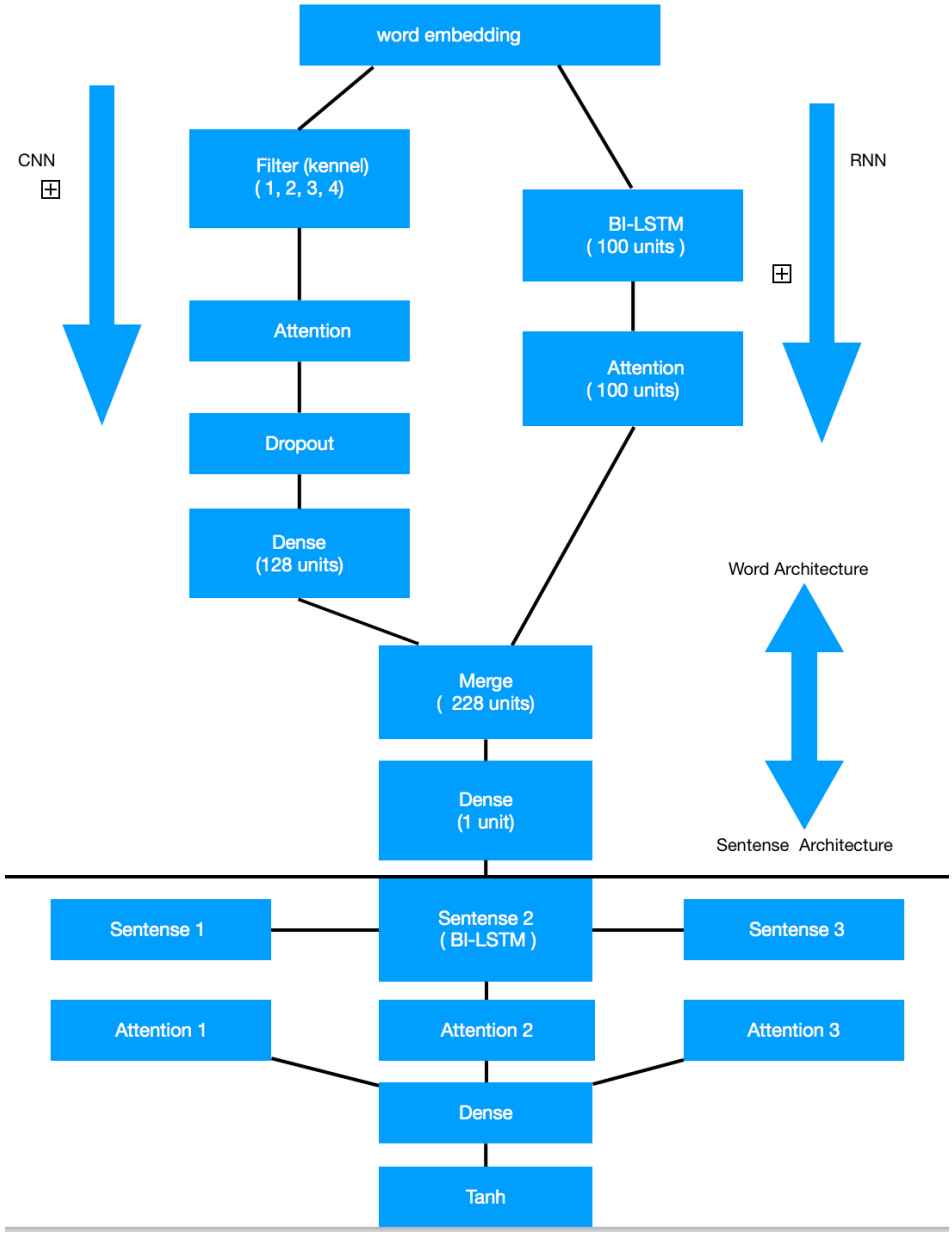
Hierarchical Attention Networks



Reference

[1] Hierarchical Attention Networks for Document Classification

Neural Network Architecture



Reference

[1] Sentiment Analysis on Financial Data Using Neural Networks

[2] My gitHub example attention CNN and RNN for sentiment analysis

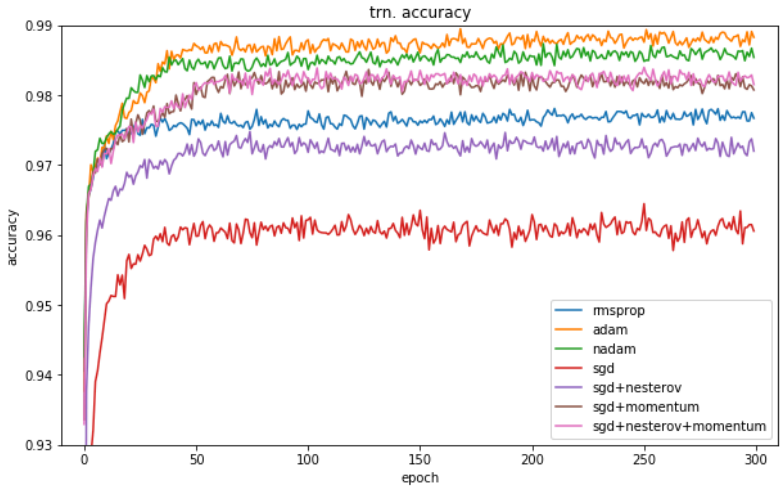
Gradient Decent

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t.$$

We compute the decaying averages of past and past squared gradients respectively:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

default values of 0.9 for β_1, 0.999 for β_2, and 10^{-8} for ϵ



Reference

[1] SGD > Adam?? Which One Is The Best Optimizer: Dogs-VS-Cats Toy Experiment

[2] The Marginal Value of Adaptive Gradient Methods in Machine Learning

[3] An overview of gradient descent optimization algorithms

Training and Metrics

Training datasets:

StockTwits, Yahoo Finance, from SemEval-2017 Task 5

Accuracy: about 70 %

My NLP Interest

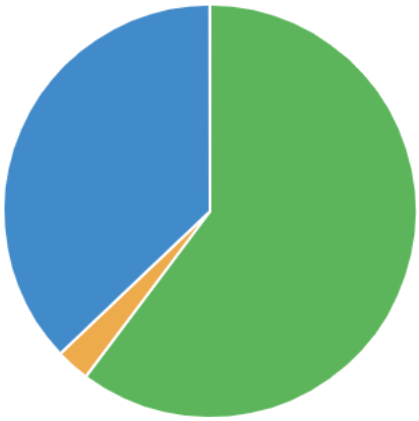
- Sentiment analysis
- Text similarity
- Document recommender system
- NER
- Topic modeling

Question ?

Coding Ability

- Python
- C/C++
- R
- MySQL/PostgreSQL

Leetcode



323
Todo

525 / 871
Solved

23
Attempted

Exp. 356 | 356