UNIVERSITY OF TRENTO - Italy
Information Engineering
and Computer Science Department

Master Degree in Computer Science

Natural Language Processing

AA 2015-2016

Sentence compressor

Aliaksandr Siarohin

September 17, 2016

Abstract

This is a description of project on natural language processing.

1 Introduction to the designed system

In this project I impliment an neural network solution to deletion-based sentence compression where the task is to translate a sentence into a sequence of zeros and ones, corresponding to token deletion decisions.

2 The analytical model

2.1 Pre-processing

The data was initially in json format. To parse json I use script provided by teaching assistant. After parsing the data become the list of list with tokens. Each token is tuple with 3 field (word, tag, stem, token deletion dessision). I populate this tuple with additional field word embeding. This embeding is taken from google-news word2vec `https://code.google.com/archive/p/word2vec/`.

2.2 Formal description of neural network architecture

I use LSTM based network for this problem. The input to my network is word embeding and one hot encoding of word tag. Then there is 2 LSTM layers one which traverse the sentence from the begining and the other which traverse it from the end. The output of this layer go to dropout layer, and the output of dropout go to the same 2 LSTM layers. Then there is dence layer with dropout. And finally dence layer with softmax nonlinearity. The network structure can be found in fig. 1.

3 Software description

To impliment this project I use python librarys theano+lasagne. This framework can build a graph of execution, and then evaluate it on gpu.

3.1 Used Functions

I use different network layers from lasgne library (LSTMLayer, DropoutLayer, DenseLayer ...). For fiting network I use adam function from lasagne library.

4 Experiment Description

4.1 Data

Data is 10000 sentences. For each word in sentence there is stem form of this word and tag of this word. Each word has assosiated token deletion decisions (0 or 1). For the experiment I split the data in 3 sets: train, dev and test. Test set is 1000 first sentences, train and dev set is random split of the rest 9000 sentences, 1000 for dev and 8000 for train.

4.2 Training process

The network is trained using adam method. The batch in my case was 1 sentence. The sentences are feeded to network in random order. The training process takes 4 epochs. After each 1000 sentences I compute score for dev set and save network parameters. The network with the best score on validation set will be result.

4.3 Network parameters

The size of word embeding is 300 (Because in google-news dataset it 300). The number of units in LSTM layer is 200, as well as in tag embeding. The size of last dence layer is also 200. Last layer give us probability, so in order to get 0, 1 predictions we need some threshold. In my case this threshold maximizing the f1 score on dev set and it equal to 0.505582.

5 Result presentation

The result score on the test set:

- Per token accuracy: 0.832198
- Per token f1 score: 0.867230
- Fraction of right compressions 0.162000

The learing curves can be found in fig. 2. The Precision/Recall curve can be found in fig. 3.

The table with compression examples can be found in table 1 for good compressions and table 2 for bad compressions.

Table 1: Good compressions

Starting sentence	Reference	Predicted
Ashton Kutcher recently chose Angelina Jolie over Mila Kunis during a Bang Marry Kill game in an episode of the television series Two and a Half Men	Ashton Kutcher chose Angelina Jolie over Mila Kunis	Ashton Kutcher chose Angelina Jolie over Mila Kunis
Mobility is changing the way enterprises are buying deploying and using unified communications according to research firm Canalys	Mobility is changing the way enterprises are buying deploying and using unified communications	Mobility is changing the way enterprises are buying deploying and using unified communications
Jeter said that he will retire after the 2014 season according to a post on Facebook	Jeter will retire after the 2014 season	Jeter will retire after the 2014 season
A plane carrying several cadets and officers from the United States Air Force Academy in Colorado Springs made a safe emergency landing in the northern New Mexico city of Las Vegas due to ice build-up on the aircraft	A plane carrying cadets made a safe emergency landing	A plane carrying cadets made a safe emergency landing
Earlier today we reported that Glee could be BANNED from the UK after a comedy club franchise The Glee Club won a high court order against the hit US show Glee	Glee could be BANNED from the UK	Glee could be BANNED from the UK
Aktia Bank is renewing its core banking system with a completion date in 2015	Aktia Bank is renewing its core banking system	Aktia Bank is renewing its core banking system

Table 2: Bad compressions

Starting sentence	Reference	Predicted
However that means high quality players are being left on the bench so which of this Newcastle quintet should feel most hard done by	that means so which of this Newcastle quintet should feel most hard done by	that means high quality players are being left on the bench
Gulf Finance House has signed an agreement with a consortium of British investors in order to sell 75 per cent of their stake in Leeds United Football Club	Gulf Finance House has signed in order to sell 75 per cent of their stake in Leeds United Football Club	Gulf Finance House has signed an agreement with a consortium of investors
Stocks to watch on the Australian stock exchange at the close on Tuesday	Stocks to watch at the close on Tuesday	Stocks to watch on the Australian stock exchange
Stocks to watch on the Australian stock exchange at the close on Wednesday	Stocks to watch at the close on Wednesday	Stocks to watch on the Australian stock exchange
The second wild is the picture of all five of the Girls people roulette online chat Guns	people roulette online chat Guns	The second wild is the picture of all five of the Girls people roulette Guns
A US woman who shares her home with 50 skunks says they are wonderful beautiful animals	A woman who shares her home with 50 skunks says	A US woman are wonderful beautiful animals

6 Discussion

6.1 Basic considerations on the starting software and the obtained improvement (of such software)

6.2 Implementation problems and some characteristics of implemantation (e.g. computational complexity, execution time and usability).

6.3 Comparison among different presented models (explanation of the improvement or decrease in accuracy)

7 Conclusion

7.1 The main (and few) main points and results of your work

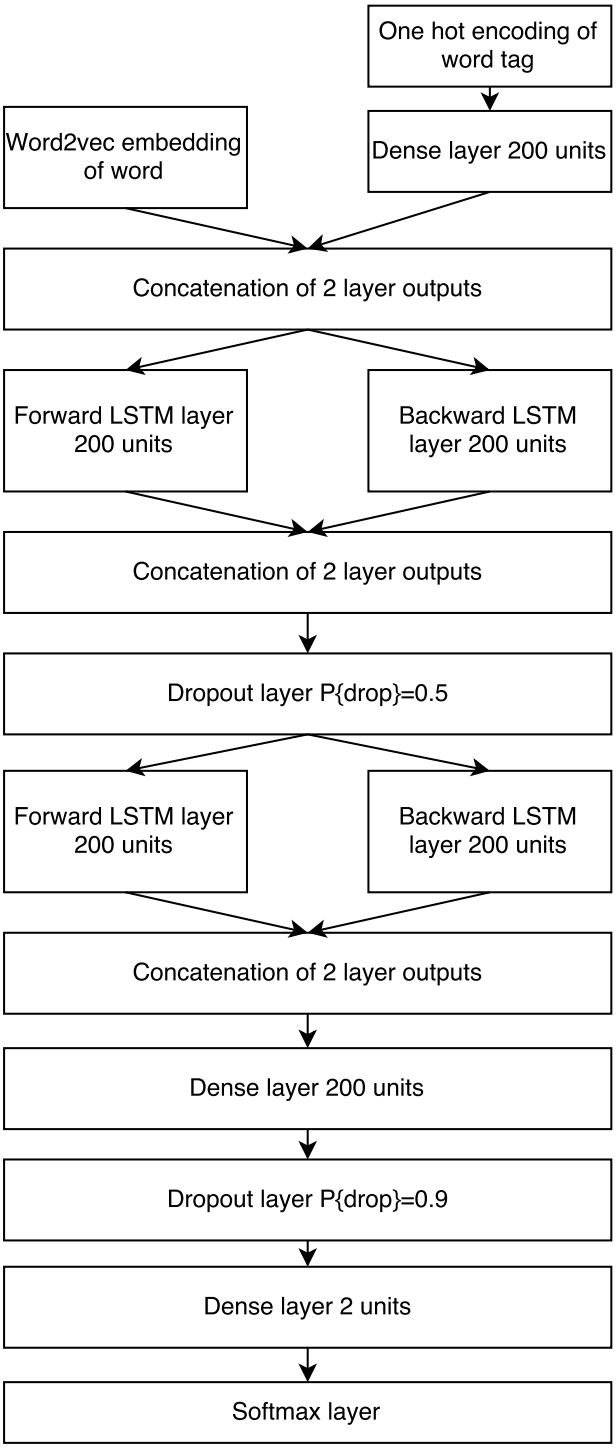

Figure 1: The network architecture

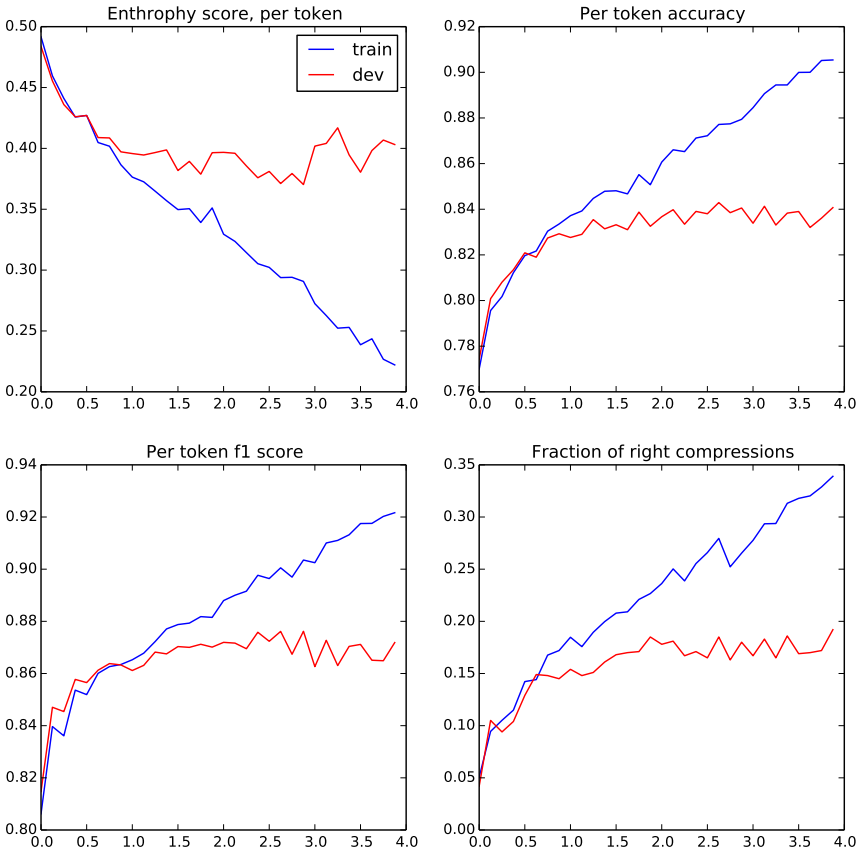

Figure 2: The learning curves on train and dev sets

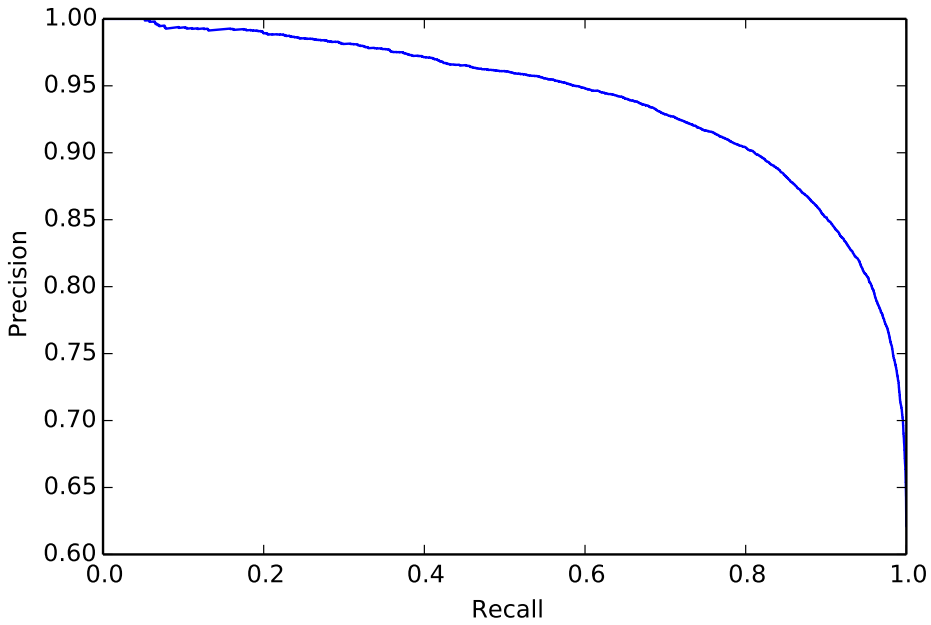

Figure 3: The Precision/Recall curve