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 18, 2016

Abstract

This is a description of project on natural language processing.

1 Introduction to the designed system

In this project I implement 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 decision). I populate this tuple with additional field word embedding. This embedding 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 embedding and one hot encoding of word tag. Then there is 2 LSTM layers one which traverse the sentence from the beginning 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 dense layer with dropout. And finally dense layer with softmax non-linearity. The network structure can be found in fig. 1.

3 Software description

To implement this project I use python library's 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 lasagne library (LSTMLayer, DropoutLayer, DenseLayer ...). For fitting 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 associated 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 feed 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 embedding is 300 (Because in google-news dataset it 300). The number of units in LSTM layer is 200, as well as in tag embedding. The size of last dense 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: 0.774660
- Fraction of right compression 0.162000

The learning 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 compression and table 2 for bad compression.

Table 1: Good compression

Starting sentence	Reference	Predicted
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
The Galaxy S III is finally receiving the Android 4.3 update in India more than a month after the update started rolling out in the US and a few other regions	The Galaxy S III is receiving the Android 4.3 update in India	The Galaxy S III is receiving the Android 4.3 update in India
Nice Ride bikes will make their spring debut on city streets Saturday and riders can now keep them for twice as long	Nice Ride bikes will make their spring debut	Nice Ride bikes will make their spring debut
Not following its key ally Saudi Arabia 's on Syria Pakistan Thursday asserted that Syrian people should decide their future set up	Syrian people should decide their future set up	Syrian people should decide their future set up

Table 2: Bad compression

Starting sentence	Reference	Predicted
A MAN who was shot in the head in front of his teammates as he walked off a soccer pitch knew his life was in danger an inquest heard	who was shot in the head in front of his teammates as he walked off a soccer pitch	A MAN knew life was in danger
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	players are being left
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
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

6 Conclusion

In this project I implement an neural network solution to deletion-based sentence compression. Resulting network show good results on phrases with one subject and predicate. But not very good at phrases without subject, as well as complex phrases with 2 or more subjects.

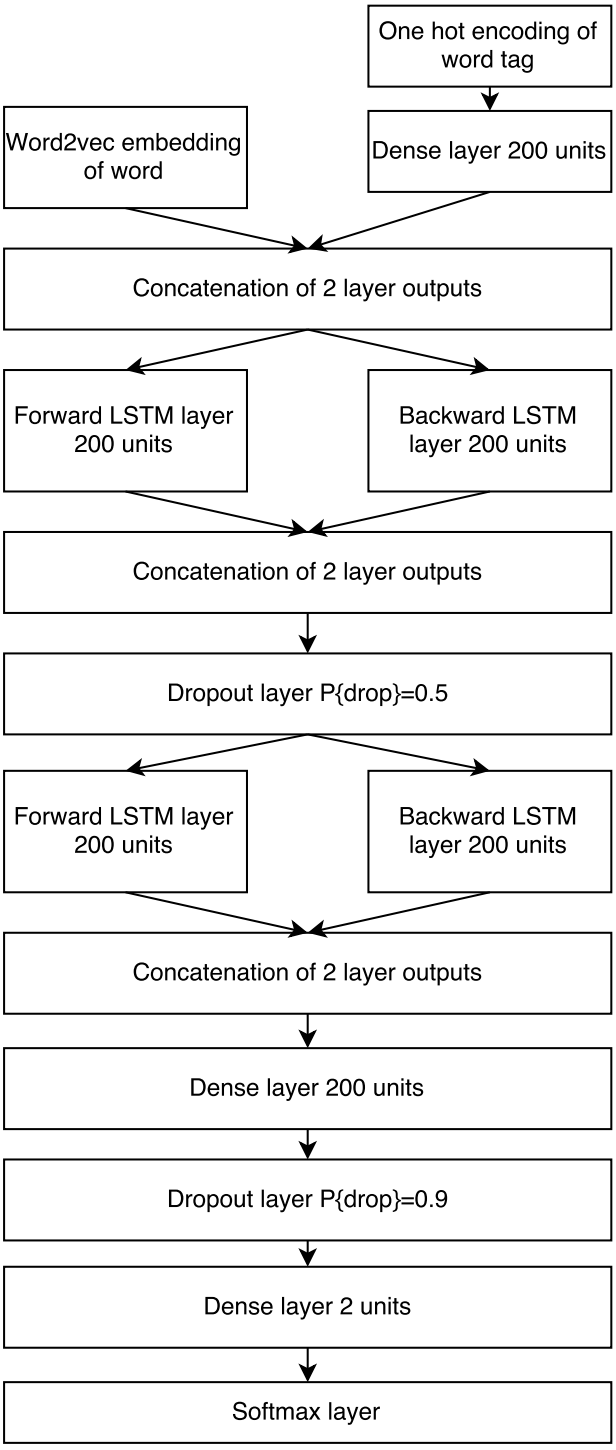

Figure 1: The network architecture

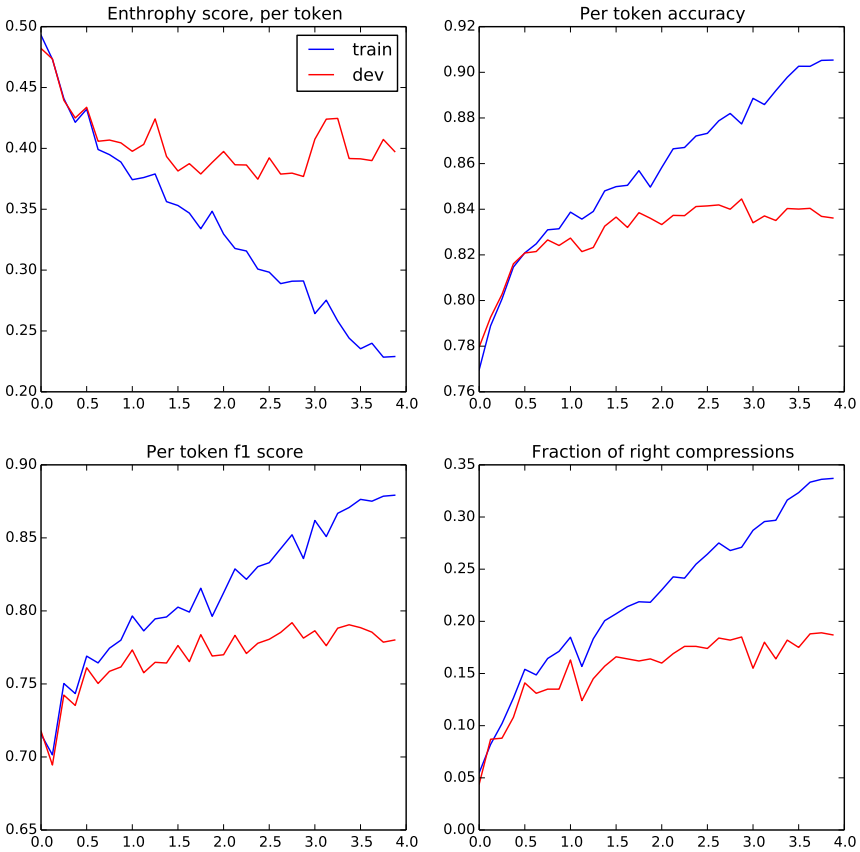

Figure 2: The learning curves on train and dev sets

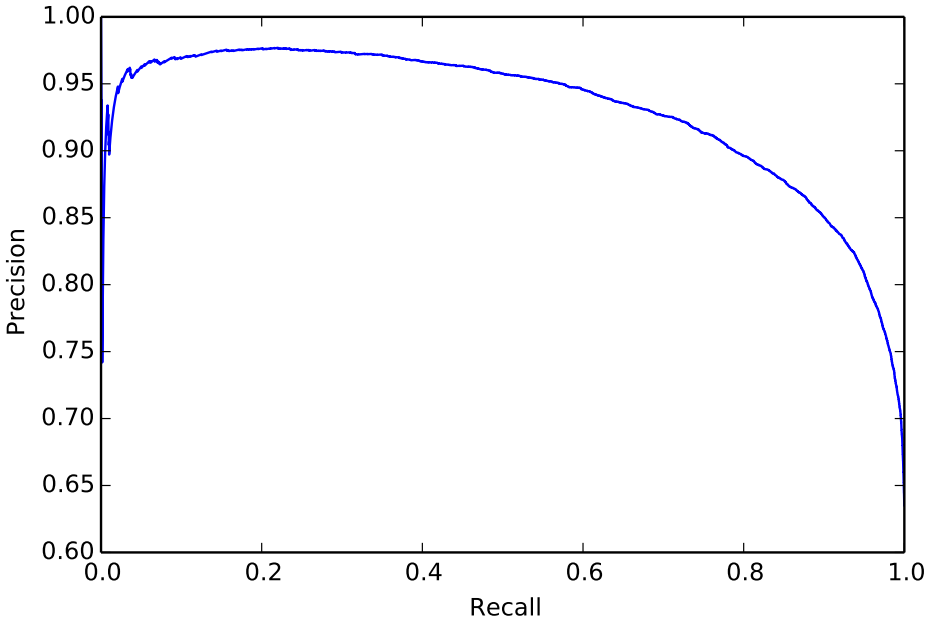

Figure 3: The Precision/Recall curve