Blue – Interviewer1, Purple – Interviewer2, Green - Interviewee

So, thanks for agreeing to do this. So, I think we will start by explaining you what is the objective of this interview. So, we are conducting a study and our aim is to create a taxonomy of the problems that developers encounter when they are developing systems that use deep learning and machine learning. So, it is like in different parts and one part is that we analyse GitHub commits and issues and StackOverflow discussions on the frameworks that are related to deep learning, such as Tensorflow, Keras and Pytorch. But we think that you can find only specific types of bugs and problems in this setting, like we do not know how people collected their data, how they were tuning their model. It is mostly like more kinds of bugs that give you an error message, while we think that space of errors is much bigger than that, so that is why we are conducting these interviews with the developers to get a wider picture on the real-world problems in deep learning systems.

Just to confirm that you have agreed that we will tape this call and that we will share the transcribed version of it, but all the data will be anonymous.

So, at first we want to ask some questions on your background.

What is your current position?

I am working towards finishing my PhD. And I am working in Computer Science domain, more specifically I am working on Natural Language Processing problems. This includes different methods such as statistical machine learning algorithms, but I have experience of working with deep learning models as well.

So, how much of overall work experience do you have?

If we consider only industrial experience that would be 2 years and 9 months, but I have been a PhD student for almost 4 years now.

Thank you. So, for how long approximately you have worked with Deep Learning/Machine Learning systems?

I would say overall maybe for 3 years, 2.5 years.

Can you give us some specifics on what type of DL/ML networks have you developed/implemented? And by that I mean supervised/unsupervised/reinforcement learning or any more specific types that you can tell us.

So, for my task I worked mostly on supervised machine learning neural networks. Given the nature of the task, these are mainly LSTM networks, which is long short term memory networks. They help to capture long-term dependencies in text, such as sentences and so on.

Ok. Thank you. Which problems were you trying to tackle using the deep learning network? Like image classification, speech recognition, this type of problems.

It is called tagging problem, but I was not tagging independent entities, but sequential data. I was trying to assign tags in a sequence.

But what kind of tags? Like you have a sentence, what you should do with that?

So, my sentences are arguments and I am trying to identify the internal structure of those arguments, meaning what is the premise, where does the boundary of the premise start, where does it end, where does the conclusion start and what are the relations or links between those two parts of an argument.

Ok. Thank you. And the last general question is which programming languages and frameworks have you been using in the process?

I am mainly using Python and Tensorflow.

Ok, from here we could go two ways maybe. Our main question is what kind of problems, like bugs and problems for achieving the desired level of accuracy and performance you have faced when you were working with deep learning systems. Do you have anything on mind that has happened quite often to you?

There are different kinds of problems. I am not sure how to classify them.

Yeah, if you could classify and tell us as much as you remember that would be very useful.

I think one of the main problems that I had has to do that the fact that the messages that I have while developing my models are not very informative. I am working on a Windows environment, and as we know Tensorflow is initially developed for Unix machines, so there are certain problem. I am working on Unix environment for Windows to develop most of my models and I am using Anaconda IDE. So most of the time the errors that I have are not very informative and it takes me a lot of time to understand where the error actually happened.

So, you mean the error messages are not clear?

Yes.

Ok. But is it something specific to this Windows thing? Or is it a general Tensorflow problem?

No, I just gave the background. I think it is the general Tensorflow problem. Although, it might have to do with my IDE, cause when I run it from the command line on Windows, the message is usually much more detailed than the one I get from Anaconda.

Is there anything else you want to add?

Yes, then, we have a lot of memory problems. Again, I can talk about my machine. For example, if I develop a large enough model with a huge data input I often have “the kernel died” error message, which most of the time has to do with algorithm not being able to allocate enough memory.

So, do you think you could get away of this one if you had a more powerful machine? Is it the solution? Or do you think you have to pre-process your training data in some way to make it work? Or how do you deal with it when you get the error?

Yes, I try to allocate enough memory usually. Or I run it on my work machine, which is Unix and it has enough memory.

Have you considered dividing data in some way? Like batches or something.

Yes, yes, I sometimes have to do that.

Do you have anything else to add? Or should I ask a question?

You can ask something.

I have one question. You say that the messages are not informative enough when you use Anaconda. So, usually, you find the solutions to these problems eventually. And maybe, you could say like which problems were the most prevalent, problematic to analyse.

Well, as I said I have this message, “the kernel has unexpectedly died” and it does not give me a lot of information, so.

So, it was all about memory?

No, not always. It sometimes returned me wrong lines of where the error has occurred. And I was really searching for the error on that line, like a specific procedure, but it turned out that the error has actually happened a couple of lines earlier, as I started debugging and analysing it.

But, these errors are usually memory stuff?

Part 2.

-Sorry for the interruption, we had a connection problem.

- No, actually...

-So where did we stop? You were telling us that it is not always the memory problems and that you do get something else.

- Yes, like sometimes it has to do with the wrong dimensionality of my data, but I cannot immediately see where the problem is because I get just the wrong line for my error.

-Okay. Anything else?

- It takes a lot of time to tune my hyperparameters often, but I think it is a general problem with most of the ML networks

-So how do you decide on the values of hyperparameters?

-It depends on the certain hyperparameter. I usually select the initial one as a standard that is used in literature and then I just increase or decrease it depending on the task and see whether it improves.

-So, can you name like a specific hyperparameter that causes your problems most of the time?

-It is usually the learning rate, for example.

-Okay, so maybe I can ask you a question. So, when you train your networks do you use some existing datasets, or have you actually ever collected your own data?

-So, in my task, I, unfortunately, do not have any developed datasets, so all of my data have been manually (not manually, it was collected automatically), but annotated manually for supervised learning.

-Okay, how did you collect it automatically?

-Um, so part of my data comes from British National Corpus, part of it comes from Gigaword which is a corpus of news articles and approximately 1/3 of my data is collected from social media, more specifically, from Twitter, using Twitter ad on for Python. So, I just search for specific twits and extract those twits.

-So, you mention that you do the manual labeling.

-Yes

-Have you done it yourself or have you ever had other people doing it?

-So, for my training set I usually annotate it myself. But my test set is something that was annotated by independent two annotators and it consists of the sentences on which both of the annotators agreed on specific tag.

-Okay, so, do you have some kind of check on this annotation? So, if you annotate yourself, you just believe your own judgement if I understand correctly, right?

-That’s true

-So, yes, and for the test set, when you had two annotators, have you cross-checked them or have they been tagging different things like different data or maybe the same and you did the cross check?

-Yes, so for mine, unfortunately, I do not have anyone else to judge on my annotation, because it is costly, like there exist some financial limitations to this study, but for the test set, I only select those on which both of them agrees, so they annotate the same thing and I always select those that \*inaudible\*

-Have you tried to use a test set, like separate one from each of them? And then only the one on that they agree on.. Did you see any difference in accuracy doing that or you did not try that?

-Sorry, I am not sure I understood you.

-If you have two annotators, you have two different test sets, right?

-No! It is the same, it is the same test set.

-It is exactly the same?

-Yes, annotated by the \*inaudible\*

- \*inaudible\* So you have from the beginning just took the one they agreed on?

-Yes

-So, do you have a number which was the rate of agreement? Do you remember it?

-It was 75, 75 with something, but I do not remember exactly.

-Okay. And there was a case when you gave annotation for your training data and mb you got something wrong like a low accuracy and then you changed your annotation like, ‘corrected’ yourself in a way?

-So, I am not sure what do you mean by accuracy. Is it the accuracy of my network or my own accuracy?

- The network, yes. So, maybe like you do the training, you see that it is not working properly, and you think: “Okay, I did not do the right annotation and that is the reason”.

-Sorry, It has frozen so I couldn’t hear what you said before.

-I was telling that it is the accuracy of the network and maybe, I am just guessing, but maybe you did train on your initial annotation, your model did not perform well, and you went back to it to correct your annotation? Did you ever have any kind of problems like that?

-No, I usually do not return to, like do not remark my annotations.

-Okay, thank you. So, in general, what kinds of problems related to training data do you face in the progress? You mentioned that it is expensive for you (Uhum), but do you have to deal with it in any other way like ordering, shuffling or thing like that. Or do you do any preprocessing for the collected data?

-So, it depends on the task, but for some tasks… so, I am trying to train it currently as a complete end-to-end model and it is like the networks that are stacked on the top of each other so one of the problems is how to connect one layer to another like deal with the dimensionality, for example. Another one is… But for some tasks I do need to make a lot of preprocessing in terms of for example, post-tagging this information, assigning partial speech tags for every word, then cleaning my data – removing all the stopwords. Sometimes performing the structural analysis of the sentence, so there could be a lot of preprocessing in this.

-Can you give us an example of a problem you faced because you did a wrong preprocessing on something?

-The wrong preprocessing?

-Yes, or you did not do it at all and then you thought that there was a need for it?

-Well, I think that the main problem during preprocessing could be the fact that the preprocessing tools themselves are not working ideally. So, often I have wrong tags assigned to certain words and this of course results in wrong interpretation of that data, assignment of wrong tags, and when I look back to features that are produced after preprocessing, it turns out that tools themselves , I use Stanford Parser for example, are not ideal.

-Okay, so, can you tell us more about the structure of the models that you have been using? SO, do you usually have to decide on the type of the layers and etc., or do you use some predefined models for your task?

-I think the main idea behind my models is by default predefined, like I do not invent my own types of networks, but the architecture and which exactly type of model I apply, what kind of input I provide – I model it myself. So, to provide some information on an input, I first have pretrained word embeddings as a data. So, where I have like already collected embeddings for each word – they were pretrained on Twitter, they are called ‘Glove’ embeddings and I then take like a separate data from the same corpus, so it is also from Gigaword, but these are not the arguments I am using for my annotation task. And what I do, I train them for character embeddings, and I use convolutional NN for this task. So, by doing this I can have certain morphological features for each word, which are important when I, for example, try to define the negation in a sentence or in a word. And what I do then is I just concatenate the two representations - the character embeddings and my word embeddings, and this is the input to my network. And then, on the top of that, the model itself consists of long short-term memory networks, but for my task, like the specifics of my data is that I need to have access to feature context in order to annotate the current context, so I cannot use because the way long short-term memory work is they predict the next input based on what they have seen, but I need to have access to feature information because I am doing annotation and my data is parallel in a way, so I need to see the feature tag in order to understand the current tag. So, I am using the bi-directional LSTM instead of just normal LSTM which means that on the top of the normal LSTM I provide…I input the same sentence but backwards and then I concatenate the final representations to have the full representation.

-So, I would like to ask about the problems that you have faced when trying to create these structures, these layers or like did you have any problems \*inaudible\* number or the type of the layers?

-Mm...

-So, because what you have told us so far is about the input data, and then you are passing it thorough the LSTM, right?

-Yes

-So, mb you did not ever have this kind of problems?

-Maybe they were technical, and I just do not remember what exactly the problems were. I tis like usually dimensionality reduction or selecting proper features because the embeddings can get very large and you have the embedding for each word, so the feature selection is an important thing. In my convolutional NN there is the pooling algorithm that kind of aggregates certain characters together so just deciding how much of the information you want to aggregate as an output is something that you achieve after many trials and so on.

- So, I wanted just to quickly go back to the training data. I wanted to ask how you decide if you have enough training data or if you had problems because you did not have enough training data.

-To be honest, if you ask me now, I would say I do not have enough training data because for the similar tagging tasks the number of sentences people usually have is hundreds of thousands, about 200 000 sentences and my dataset is roughly 4000 sentences so it is about 20 times less than usual cases.

-Not recorded

-No, I use the L2 loss function which is the standard type of the loss function for many NLP tasks.

-Not recorded

-No, No

-Not recorded

-Like, if I remember correctly, it was about the fact that I was multiplying matrices and I had to transpose them because you have to specify the dimensionality for an output and it just did not match the actual dimensionality that I tried to assign to it.

-Not recorded

-No, but just in general, when you work with these frameworks such as TensorFlow, which is a very handy tool, helps to solve a lot of ML problems, but one of the problems that I have is when you use a function, it is often not very obvious what is happening on the lower level -so how exactly is this function calculated, is it just that you input something into certain type of, I do not know, training or model and it is hard to understand sometimes what it is doing in the lower level.

-Not recorded

-I can actually be more specific, let me check now. Okay, so, in my case, because it is not like one type per input. So, I have the chains and they can have boundaries – I am not measuring the accuracy, but I have precision recall and F1 score which is the harmonic mean and like with the final model I get F1 of 83 on the development side and 82.7 on the test set.

-Not recorded

-Well, one of the steps is definitely data collection and increasing my training set, then tuning hyperparameter of course, then another step was reducing the number of labels because initially we had about 15 label of final tags and then we merged them into 7 final tags.

-Not recorded

-Yes, because even on a semantic level they were not mutually exclusive, so we merged some of them.

-Not recorded

-Just identifying how many epochs do you need to get better results is also an important thing because after just two epochs the results were not great.

-Not recorded

-So, what I usually do is I specify a high enough number, but then somewhere in my I have a check: if after, for example, 6 epochs the results are not changing at all then I just stop.

-Not recorded

-The most ridiculous, probably, was when I had to spend a couple of days trying to solve something that was not something that actually cause the problem, the reason was an uninformative error message, but I do not remember what exactly it was about.

-Not recorded

-Thank you too, sorry for the technical issues, bye.

List of the most annoying errors:

1. Cannot read byte at position X. It happens the most when the data is from social media and people use different encodings, not just utf-8. The error can come out in the middle of the training and it is hard to fix it because the decode also do not always work. There is no universal decoding when the whole data is in different encodings and contains special symbols.
2. The second annoying error is when some packages/modules are removed from Python3. For example, the Cpickle.
3. The third one is purely Windows related. When you run Theano, some of the modules it calls like Lazylinker and etc. do not work through the Anaconda because there are problems with mingw path environment, and you have to set everything manually and it takes a long time.