Am Geisenberg 4
66125 Saarbrucken
Germany
☎ +49 17687136243
✉ deb_4u22@yahoo.com
debjitpaul.github.io/debjit.github.io

Debjit Paul

Experience

Jan.2018–Current **Data Scientist**, *Amplexor International*, Germany.
- **Department:** Content Intelligence Team
- Machine Learning Concept - Working with Neural Machine Translation (open NMT),
- NLP Concept - Working with Language Modelling toolkits (KenLM), Moses Toolkit (Tokenization, Normalization, True-Casing).
- Handling Big Data - Data Acquisition, Data pre-processing , Web Crawling.
- Language - Python.
- Development tools - Eclipse (pydev), Jenkins, pytorch , tensor-flow, SVN, Amazon Web Server.

Oct.2016–Nov.2016 **Research Assistant**, *Saarland Univesity*, Saarland, Germany.
- **Department:** Foundation of Exact Algorithm
- Programming tasks for the organization of the programming challenge **PACE**.
- The goal is to generate instances to be used in the challenge.
- **Technology:** Python, Shell Scripting **Github Page**

Nov.2015–Feb.2016 **Research Assistant**, *Saarland Univesity*, Saarland, Germany.
- **Department:** Machine Learning Group
- Crafted Algorithms using Python to filter data (Arxiv dataset) collection results.
- Concept utilized are Natural Language Processing, Prepossessing text data.
- **Technology:** Python, Shell Scripting.

April 2015–July 2015 **Student Research Assistant**, *Saarland University, DFKI*, Saarland, Germany.
- **Department:** Information and Services Systems
- Developing features in OntoUML.
- Designing Information Systems based on Conceptual Modeling.
- **Technology:** Java, Protege, OntoUML.

Education

2014–2017 **MSc. Computer Science**, *Saarland University*, Saarbrücken,Germany, *Grade-1.9 on German Scale*.

Thesis Multitasking Learning With Unreliable Labels

Supervisor Prof.Dr.Dietrich Klakow, Head of Language Spoken System Group

Description Neural networks have shown impressive performance for various classification tasks due to the availability of large datasets. However, a known problem in classification tasks is the presence of unreliable labels. Successive manual annotations can improve the quality of the data set but are expensive to obtain. Hence, most applications rely on the artificial annotation (e.g., Brill tagger for POS tags) of large data. We build a Multi-Task Learning-based (MTL) Noisy Label Neural Network. In this work, we establish Chunking as the primary task with POS tag classification as the auxiliary task. We set up the data by adding artificially annotated data with the human annotated Penn Treebank data. In our experiments, we compare the MTL-based setup with and without noise reduction against NLNN on the single task. We show that MTL-based NLNN outperforms the state-of-the-art for Chunking. **Github Page**

2010–2014 **B.Tech in Computer Science**, *GuruNanak Institute of Technology*, Kolkata,India, *Grade-8.73/10*.

Thesis Improved Algorithm for Human and non Human Object Detection

Publication

- Debjit Paul, Mittul Singh, Dietrich Klakow, *"Handling Noise in Automatically-Annotated Data with Noisy Label and Deep Multi-Task Learning Neural Networks"*, Short Paper on NAACL-HLT 2018 (Under Review)
- Anup Kumar Thander, Goutam Mandal, Debjit Paul *"Numerical Comparison of multi-step iterative methods for finding roots of non-linear equations, International Journal of Mathematics Trends and Technology "*, Volume4 Issue 8-September 2013 [ISSN: 2231-5373]

Selected Project

Topic DeepNets-onWikiData

Description This project has basic implementations of CNN, RNN (vanilla and LSTM), and various combinations thereof. We use Keras with TensolFlow backend. We use GloVe vectors for the word embeddings, which can be downloaded from here: **Glove**. **Github Page**

Topic Pre-Processing and NLP Tagger Tool

Description Transforming a unstructured text data into CONLL format NLP tagging file. This tool can handle large unstructured data. Then can be used to tag words using artificial taggers such as Senna or Brill Tagger. **Github Page**

Topic Bidirectional long short-term memory

Description In recent times Bidirectional long short-term memory (BiLSTMs) networks has proven success for several NLP sequence task such as POS-tagging, NER tagging and Chunking. Implemented a BiLSTMs tagger with word and unicode byte embeddings.

Technical skills

Language Python, R, Shell Scripting,Java, C,Matlab, MySQL

Concepts Machine Learning , Statistical Learning, Data Mining

Tools Pycharm, word2vec, Tensorflow, Keras, sckit-learn, Dynet, Standford CoreNLP, Theano, Eclipse, Matlab, Numpy

Achievements

- Winner of HQ Hackathon 2017, at Trivago

Languages

English	**Proficient**
German	**Basic**
Bengali & Hindi	**Native**

References

- Available upon request

Declaration

- I hereby declare that all the details furnished above are true to the best of my knowledge and belief