ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

Scientific Report

First name / Family name	Deepanwita Datta
Nationality	Indian
Name of the *Host Organisation*	NTNU, Norway
First Name / family name of the *Scientific Coordinator*	Heri Ramampiaro
Period of the fellowship	01/05/2019 to 31/05/2020

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Electronic patient records (EPR) are valuable text sources, where almost all the decision-making processes of medical staff are written. Thus, for acquisition of decision making and diagnosis, text mining of EPRs is very important. The discharge summaries, which include the compact explanation of medical history and treatments throughout patient's admission period are also very informative and could be used to train the learning models which could aid in diagnosis. With the progress in machine learning, extracting valuable information from medical literature has gained popularity among researchers. Information extraction or retrieval from medical text data relies heavily on the representation models. However, in medical text-mining tasks, most popular representation model, such as Word2vec, cannot be applied directly on medical data because of significant differences in vocabulary and expressions between a medical corpus and a general domain corpus. Hence, there is a need for better representation models for the medical corpus.

We began with the idea of retrieving relevant information or data as per the clinician's needs from the medical repository with the central focus being Cerebral Palsy in infants. Dealing with such sensitive cases like child health, promptly and quickly, makes the clinicians stressed out all the time. Moreover, the ever-growing number of patients to attend is making the situation more challenging for the medical professionals. So, consulting any applicable topic and searching it manually is infeasible for them. It would be beneficial for the clinicians if they can get a quick access to the particular topic automatically and the

exact relevant information they are looking for. However due to lack of availability of any open source data, we had to resort to building our own collection. Since the data is sensitive and prone to privacy and security concerns it had to be ratified by several authorities including European Commission, Ministry of Health and Care Service, Norway along with NTNU and St. Olav's hospital. This was a time consuming and slow process. In the meantime, while the data was being collected by other researchers, I chose to focus on a similar research problem with a readily available open-source data on Precision Medicine.

Precision medicine is a medical paradigm in which treatments are customized entirely to the individual patient. The underlying issue that drives precision medicine is that for many complex diseases, there are no "one size fits all" solutions for patients with a particular diagnosis. The proper treatment for a patient depends upon genetic, environmental, and lifestyle choices. The ability to personalize treatment in a scientifically rigorous manner based on these factors is thus the hallmark of the emerging precision medicine paradigm.

A fundamental difficulty with putting the findings of precision medicine into practice is that–by its very nature–precision medicine creates a very large space of treatment options (Frey et al., 2016). These can easily overwhelm clinicians attempting to stay up-to-date with the latest findings, and can easily inhibit a clinician's attempts to determine the best possible treatment for a particular patient. However, the ability to quickly locate relevant evidence is the hallmark of information retrieval (IR). For three consecutive years the TREC Clinical Decision Support (CDS) track sought to evaluate IR 1 systems that provide medical evidence at the point-of-care. The TREC Precision Medicine track, then, was launched to specialize the CDS track to the needs of precision medicine so IR systems can focus on this important issue. The Precision Medicine track has focused on a single field, oncology, for a specific use case, genetic mutations of cancer. This started with the TREC 2017 Precision Medicine track, continued in 2018, and further continued in 2019. As described above, main idea behind precision medicine is to use detailed patient information (largely genomic information in most current research) to identify the most effective treatments. Improving patient care in precision oncology then requires both (a) a mechanism to locate the latest research relevant to a patient, and (b) a fallback mechanism to locate the most relevant clinical trials when the latest techniques prove ineffective for a patient. In the first part, the track focuses on Clinical Decision Support track while in the second part expands the task to cover a new type of data (clinical trial descriptions). The main change between the 2017-2018 tracks and the 2019 track was to add the optional sub-task of determining the actual treatments described in literature articles (no changes were made for trials, where this data is more clearly available in semi-structured form). The idea behind this addition was to allow for an aspect-based retrieval approach, where results can be grouped by the actual treatments described for easier presentation for oncologists.

Most work on precision medicine focuses on developing new treatments based on an individual's genetic, environmental, and lifestyle profile. The result is a data-driven approach investigating the best treatment for an individual patient. This promising approach has led to significant advances, including an explosion of scientific research, as embodied by the Precision Medicine Initiative (PMI). This presents an information problem for clinicians, however, as the vast literature available for precision medicine can make it difficult to find the most appropriate treatment for the clinician's current patient. The ability to quickly locate relevant information for a current patient using information

retrieval (IR) has the potential to be an important tool for helping clinicians find the most up-to-date evidence-based treatment for their patients.

The document collection provided by TREC has approximately 26 million documents. Even with efficient searching mechanisms it is nearly impossible to locate the relevant document within a reasonable timeframe. To address this issue, we began by indexing the documents with an ad-hoc information retrieval system, Galago. We have experimented with several IR models including BM25, DFR (Divergence from Randomness) and Query Likelihood (Language Models). Of these, on the given dataset, DFR showed the best performance (in terms of Precision, NDCG), after empirical parameter tuning. The result of these IR models is a ranked list of top-K documents. The next step was to re-rank the selected top-K documents with natural language models such as BERT. This kind of methodology has proven successful on other datasets and tasks such as Paragraph/Passage Retrieval, Question Answering and other NLP problems. BERT models in general has to be fine-tuned for downstream tasks such as Sentence Classification, Sequence Prediction etc. In our case, we intend to infer the embeddings for each medical abstract (along with additional data such as chemicals and MESH) and query separately, and then use a similarity matching algorithm to predict the closest matches and rearrange the existing ranked list of retrieved documents. Since BERT pre-trained models are trained on Wikipedia documents, they are not the best choice for handling specialised documents such as medical or legal documents. To this end, Bio-BERT was proposed where the model had been trained on over 1 million medical documents and articles. We intend to employ the same on our selected set of documents. We would compare our results against the state-of-the-art and hope to surpass it by a significant margin.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

- Datta D, Chakraborty M, Biswas A. Your Click Matters: Enhancing Click-based Image Retrieval performance through Collaborative Filtering. In: Proceedings of the 4th edition of the Swiss Text Analytics Conference, SwissText 2019, Winterthur, Switzerland, June 18-19, 2019. CEUR-WS.org; 2019.
  Abstract:
  Image retrieval has been an active research area since the early days of computing. While ensemble, multimodal and hybrid methods coupled with machine learning has seen an upward surge replacing unimodal, heuristic-based methods; a rather new offshoot has been to identify new features associated with images on the web. One such feature is the ‘click count’ based on the clicks an image or its corresponding text gets in response to a query. Previous state-of-the-art methods have tried to exploit this feature by using its raw count and machine learning. In this paper, we build on this idea and propose a new collaborative filtering-based technique to employ the click-log of users from the web to
  better identify and associate images in response to either a text or an image query. Experiments performed on a large scale publicly available standard dataset having genuine click logs from actual users corroborate the efficacy and significant increase in efficiency of our approach.

- Under Preparation: *Let BERT be thy medicine*: A Neural Language Model for Effective Medical Documents Retrieval. To be submitted.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- Attended several seminars organised by the Data and Artificial Intelligence (DART) group at the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway.
- Attended and presented a full research paper at 4th Swiss Text Analytics Conference at Winterthur, Switzerland, 18-19 June 2019.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Place: VTT, Espoo, Finland
Duration: 1 March to 11 March 2020
Host: Dr. Arho Suominen

During the fellowship, I had the opportunity to visit the Innovation Lab of at VTT, Espoo, Finland from 1st to 11th March 2020. I was hosted by Dr. Arho Suominen. It was a fruitful visit and an enriching experience. I worked closely with Dr. Arash Hajikhani with guidance from Dr. Suominen. I presented my previous and current research work to their group which sparked immense interest among them. In the first week I met several researchers from various domain who were working with Artificial Intelligence and Machine Learning topics. They were particularly interested in working with large datasets applying topic models and latest neural network models. In particular, we were trying to find ways to employ classical topical models such as LDA and natural language models like BERT for patent clustering from a large collection of patents. I also received valuable feedback from the group to improve and extend my ongoing research work. Overall, we focused the short visit on finding common ground for future research. I felt that we had identified several prospective avenues for future research, and I look forward to collaborating with Dr. Suominen and other members of the Innovation Lab in the future.