# Approaches to abstractive summarisation

Knowledge Lab

Birkbeck UNIVERSITY OF LONDON

**Research Student**
**Manni Singh**

**Supervisors**
**Mark Levene**
**David Weston**

Birkbeck UNIVERSITY OF LONDON

**Department of Computer Science and Information Systems**

## Research Aims

The aim of my Ph.D. is to develop effective abstractive summarization approach. In this research, we will be tackling the problem of abstractive summarisation using neural network architecture that learns the appropriate sentence creation using supervised learning. This hopefully will reduce the overhead of using external linguistic resources to generate grammatically correct sequences. An application of this type of system is news article generation where news articles could be automatically created using social information distributions such as twitter feeds, blogs, op-eds etc. This work will utilize neural network based model with the aim of learning semantic relationships instead of just long/short dependencies.

## Topic introduction

Automated text summarisation systems generate summaries having most relevant information from a document in a concise form [1]. The resultant summary could be viewed as extractive or abstractive. Extractive summarisation depends on extracting relevant information from a document. On the other hand, abstractive summarisation depends heavily on semantics and language generation. It generates the summary containing new words and phrases which are not necessarily present in the original document. For this reason, it has been considered a complex process relying on deep natural language processing [2].

## Experiments

We worked on Multilingual Web Person Name Disambiguation task [4] using NER detection. The task was about clustering the web results by disambiguating person names provide as a query: "Forename Surname". For example, there might be multiple web pages for the same name provided as query which belongs to different person sharing that name. Each query result has on average approximately 105 pages with average of approximately 20 clusters. One special cluster for each query result is named, NR (Not Related), it is used for unrelated web-pages. NER approaches use feature identification [3] for entity extraction such as capital words and context features. However, we ignored capital word extraction as the data are in the form of informal web-pages and contained multilingual texts. We started with using Word2Vec similarity based distribution to expand web-pages. The expansion was performed using top-k nearest neighbours of words.

In the second experiment, we introduced seeding by picking labels from gazetteers provided by NLTK package. We combined uni-grams, bi-gram, and tri-grams into a sequence to feed to Word2Vec. However, in this setting the batch/context sequence of Word2Vec was changed to the batch/label sequence. Our hypothesis was that labels as contexts will behave as pivots for clustering named entities. This gave us all the same class labelled words in same cluster. Further, co-training was applied based on contexts to give labels to unknown words. The results of this experiment were not so impressive due to the noise in the clusters that was analysed manually.

In the third experiment, we used Context2Vec by replacing words with contexts on the previous version of Word2Vec. Furthermore, we used TF scores to select top-2 context words from the unlabelled context to loosely match with labelled contexts. First, we made Context2Vec distribution. Labelled clusters with top TF (Term Frequency) score are selected from this distribution as centroids. Then, clusters are formed by selecting top-50 nearest labelled contexts based on cosine similarity from each of these centroids. Due to memory limitation, we picked these top-50 contexts form 20000 contexts instead of original size varying between 100000-150000. This gave us most unique 50 contexts for each label with reduction in accuracy.

## References

[1] Elena Lloret and Manuel Palomar. Text summarisation in progress: a literature review. Artificial Intelligence Review, 37(1):1–41, 2012.

[2] Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. Document concept lattice for text understanding and summarization. Information Processing & Management, 43(6):1643–1662, 2007.

[3] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1):3–26, 2007.

[4] nlp.uned.es/IberEval-2017/index.php/Tasks/M-WePNaD