ROZETTA TECHNOLOGY

MULTI-DOCUMENT SUMMARIZATION: HOW TO EXTRACT VALUE FROM YOUR 'HIDDEN' DATA

Interview with Scott Matthews, Chief Data Scientist and Professor Massimo Piccardi by David Myton

Continuous research and development underpin all of RoZetta Technology's operations. From postgraduate trainees to leading scientists, they are actively involved in pushing the boundaries of data science and technological excellence – to the ultimate benefit of its customers and partners.

One highly significant area of inquiry for RoZetta's scientists and researchers is focused on Multi-Document Summarization (MDS) aimed at prising out the hidden value in unstructured data – that is, in prolific amounts of information ranging across varied domains such as media, audio, text, imaging and sensors.

These large volumes of unstructured and semi-structured data present a challenge to organizations, investors and analysts: how can they quarry them quickly and efficiently to extract the intelligence within these large banks of information?

Trying to read, understand and present conclusions becomes a time consuming and virtually impossible task given the volume of information presented.

Enter RoZetta's MDS, which consolidates the relevant points of information scattered across numerous documents into a concise and readable summary that represents the entire set of documents.

"Once unstructured and structured text are mapped, we are able to generate document summaries based on the document content," explains RoZetta's Chief Data Scientist Scott Matthews.

"For example, a research firm may have large volumes of analysis and expert opinions stored in various formats that seem inaccessible. But by applying data science methodology we are able to extract and gather key information to present relevant text from a group of documents to save time and increase the value offered," he says.

"This ability to summarise content in an efficient and scalable way provides an enormous operational efficiency boost for any business currently working with large volumes of text data."

Organisations working with large volumes of data can benefit from RoZetta's MDS in the following ways:

- Reduced time to derive insights from text documents
- Capturing expert knowledge in a scalable system
- Consistency in the summaries generated - no human error
- Ability to ingest more sources without any impact on productivity, and
- More efficient use of high value resources - provides staff with more time to work on higher value opportunities.

The significance of RoZetta's work in this domain has been recognised by the Association for Computational Linguistics, which has accepted a RoZetta research paper for presentation at its 2022 conference – the most prestigious in its field.

This research paper – by Jacob Parnell, Dr Inigo Jauregi and Professor Massimo Piccardi - outlines a new way of tuning Multi-document Summarization (MDS) models, designed to improve the fluency and relevance of the generated summaries.

"This paper is being presented by Jacob at the very top conference for this kind of research," says Professor Piccardi, an expert researcher in Natural Language Processing and Machine Learning.

"Importantly, the paper will be published online in the conference proceedings – it's actually a live document that stays in a very consultative digital library, available on line and for free so it gets wide dissemination. It's great exposure."

Professor Piccardi says UTS is "extremely excited" to be involved with RoZetta's research efforts. "It's an excellent synergy," he says.

Prof. Massimo Piccardi

"RoZetta's highly focused research and operations make it a company that is competitive, plausible and attractive to a number of clients. It offers a bespoke service on top of what the other big companies produce – it is a very nice synergistic way of working alongside the big companies, not against them."

For enquiries contact us via email at **enquiries@rozettatechnology.com**

Read the full research paper, "*A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization*" at **Arxiv.**

