World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:19, No:07, 2025

Quantitative, Preservative Methodology for Review of Interview Transcripts Using Natural Language Processing

Authors : Rowan P. Martnishn

Abstract : During the execution of a National Endowment of the Arts grant, approximately 55 interviews were collected from professionals across various fields. These interviews were used to create deliverables – historical connections for creations that began as art and evolved entirely into computing technology. With dozens of hours' worth of transcripts to be analyzed by qualitative coders, a quantitative methodology was created to sift through the documents. The initial step was to both clean and format all the data. First, a basic spelling and grammar check was applied, as well as a Python script for normalized formatting which used an open-source grammatical formatter to make the data as coherent as possible. 10 documents were randomly selected to manually review, where words often incorrectly translated during the transcription were recorded and replaced throughout all other documents. Then, to remove all banter and side comments, the transcripts were spliced into paragraphs (separated by change in speaker) and all paragraphs with less than 300 characters were removed. Secondly, a keyword extractor, a form of natural language processing where significant words in a document are selected, was run on each paragraph for all interviews. Every proper noun was put into a data structure corresponding to that respective interview. From there, a Bidirectional and Auto-Regressive Transformer (B.A.R.T.) summary model was then applied to each paragraph that included any of the proper nouns selected from the interview. At this stage the information to review had been sent from about 60 hours' worth of data to 20. The data was further processed through light, manual observation – any summaries which proved to fit the criteria of the proposed deliverable were selected, as well their locations within the document. This narrowed that data down to about 5 hours' worth of processing. The qualitative researchers were then able to find 8 more connections in addition to our previous 4, exceeding our minimum quota of 3 to satisfy the grant. Major findings of the study and subsequent curation of this methodology raised a conceptual finding crucial to working with qualitative data of this magnitude. In the use of artificial intelligence there is a general trade off in a model between breadth of knowledge and specificity. If the model has too much knowledge, the user risks leaving out important data (too general). If the tool is too specific, it has not seen enough data to be useful. Thus, this methodology proposes a solution to this tradeoff. The data is never altered outside of grammatical and spelling checks. Instead, the important information is marked, creating an indicator of where the significant data is without compromising the purity of it. Secondly, the data is chunked into smaller paragraphs, giving specificity, and then cross-referenced with the keywords (allowing generalization over the whole document). This way, no data is harmed, and qualitative experts can go over the raw data instead of using highly manipulated results. Given the success in deliverable creation as well as the circumvention of this tradeoff, this methodology should stand as a model for synthesizing qualitative data while maintaining its original form.

Keywords : B.A.R.T.model, keyword extractor, natural language processing, qualitative coding
Conference Title : ICCCNT 2025 : International Conference on Computing Communications and Networking Technologies
Conference Location : Vienna, Austria
Conference Dates : July 29-30, 2025