Automatic Extraction and Analysis of Direct Speech from Texts Using Large Language Models

Dmytro Lande, Viktor Kuzminskyi

Igor Sikorsky Kyiv Polytechnic Institute

This paper proposes an approach for the automatic extraction of direct quotes from texts using large language models (LLMs) and their analysis to build semantic networks of authors and concepts. After retrieving relevant documents, LLMs are employed to extract quotes, their authors, and metadata, which are stored in a structured JSON format. Based on this data, a semantic network is constructed, which is then clustered using LLMs. The concept of a "swarm of virtual experts" is introduced for more precise extraction of key concepts. The model illustrates how authors form groups based on shared interests and discussion topics. One of the innovative aspects of the approach is the automatic generation of cluster names.

Keywords: *Quote extraction, Semantic network, Clustering, Swarm of virtual experts, Automatic text analysis, Large language models (LLM)*

Introduction

Today, a large volume of textual information contains important fragments of direct speech, including interviews, public speeches, and statements from prominent figures. Automatic extraction of these fragments is crucial for content analysis, media monitoring, and studying social trends.

The development of large language models (LLMs) has significantly transformed approaches to text data processing. Automated analysis of large text corpora, particularly the extraction of direct quotes, has become increasingly important, allowing for rapid retrieval of information from speeches, interviews, and publications. Thanks to LLMs, the process of extraction and subsequent analysis is faster, more accurate, and scalable.

In this work, we propose a methodology based on the use of large language models (LLMs) for the automatic extraction of quotes and the construction of networks of authors and key concepts. An important task following the extraction of quotes is the identification of key concepts related to the quotes and the creation of semantic networks of authors and concepts. These networks can be clustered, which helps in identifying groups of authors discussing similar topics. We propose using the "swarm of virtual experts" concept, where LLMs are queried multiple times about key concepts, each time simulating different expert roles to enhance analysis results.

The goal of the research is to develop a process for analyzing direct quotes and identifying thematic connections between different authors.

Literature Review

The problem of extracting direct quotes and their subsequent analysis has been discussed in many studies. One of the main challenges is the accurate identification of the boundaries of direct speech, especially in complex texts such as news articles, interviews, or legal documents. Some research focuses on traditional natural language processing (NLP) methods, such as using regular expressions, parsers, and semantic analyzers for quote extraction [1, 2]. However, large language models significantly simplify this task by better interpreting the context of quotes and tracking relevant segments [3, 4].

Another important area is the construction of semantic networks, where concepts and authors represent graph nodes, and the connections between them are edges. Semantic networks are used to study collaboration in science [5], media analysis [6], political communication [7], and other domains. Clustering such networks helps uncover hidden thematic structures and shared interests. Clustering methods, such as the modularity algorithm [8], are used to group graph nodes by common properties.

The proposed concept of the "swarm of virtual experts" is similar to approaches using model ensembles and multiple queries [9]. It improves accuracy

by generating a "family" of queries from different hypothetical experts, each modeling different perspectives on the same text.

Methods

Document Retrieval Stage

Initially, documents are retrieved using an information retrieval system based on specific keywords related to the chosen topic. The retrieved documents must be suitable for direct quote analysis.

Quote and Author Extraction

LLMs are employed to extract fragments of direct speech (quotes) from the text. Based on prompts, quotes, author names, publication dates, and source URLs are extracted. This data is stored in a JSON format, which includes:

```
{
   "title": "Title of the Article",
   "url": "Link to the Article",
   "date": "Publication Date",
   "citations": [
      {
         "author": "Author Name",
         "citations": [
            "Quote 1",
            "Quote 2",
            "Quote 3"
         ]
      },
      {
         "author": "Author Name 2",
         "citations": [
            "Quote 1",
            "Quote 2"
         ]
      }
   ]
}
```

Semantic Analysis and Network Construction

For each author, after the extraction of quotes, the LLM performs the extraction of key concepts that form the semantic network. The concept of a "swarm of virtual experts" is proposed—this involves multiple queries to the LLM, each simulating an expert with different roles, such as sociologist, politician, journalist, etc. This approach helps better account for context and uncover more connections between concepts.

Quote Extraction Stage

Let $D = \{d_1, d_2, ..., d_n\}$ } be the set of documents retrieved from the search system, where each document did_idi contains a set of textual data. We can consider the LLM as a function F_{LLM} that takes the text of a document and returns a set of author-quote pairs:

$$F_{LLM}(d_i) = \left\{\left(A_{ij}, C_{ij}\right)\right\}_{j=1}^{m_i}$$

where:

- A_{ij} is the author of the j-th quote in document d_i,
- C_{ij} is the text of the direct speech or quote from author A_{ij},
- m_i is the number of quotes in document d_i.

For all documents, we obtain the set of all quotes:

$$C = \bigcup_{i=1}^{n} F_{LLM}(d_i)$$

Semantic Network Construction Stage

Next, each quote C_{ij} is analyzed to extract key concepts. Let $F_{concepts}\left(C_{ij}\right)$ be the keyword extraction function that returns the set of concepts K_{ij} for each quote:

$$K_{ij} = F_{concepts}(C_{ij}) = \left\{k_1, k_2, \ldots, k_{l_{ij}}\right\}$$

where l_{ij} is the number of key concepts for quote C_{ij}.

Thus, for all quotes, a bimodal network is constructed with two types of nodes: authors A and concepts K. A node A_{ij} is connected to a concept k_p if that concept was extracted from quote C_{ij}. Formally, this can be represented as a graph $G = (V, E)$, where:

- $V = A \cup K$ is the set of nodes (authors and concepts),
- $E = \{(A_{ij}, k_p) \mid k_p \in K_{ij}\}$ is the set of edges between authors and their concepts.

Network Clustering

Modularity algorithms allow for the grouping of authors based on thematic similarity. LLMs can also perform network clustering based on modularity and autonomously determine cluster names by analyzing common concepts within them.

The modularity of the network is calculated using the formula:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where A_{ij} is the element of the adjacency matrix, k_i and k_j are the degrees of nodes i and j, m is the number of edges, and $\delta(c_i, c_j)$ is a function that equals 1 if nodes i and j belong to the same cluster.

LLM uses the results of clustering to create cluster names based on the concepts that are most frequently encountered in each cluster.

Cluster Naming Stage

For each cluster C_i that contains authors $A_1, A_2, ..., A_p$ and concepts $K_1, K_2, ..., K_q$, the LLM can generate a name based on the most significant concepts within the cluster. The selection of these concepts can be made by evaluating their weight in the cluster based on their frequency or centrality in the network.

Let the weight of concept k_j in cluster C_i be defined by the function

$$\omega(k_j, C_i) = \frac{\eta(k_j)}{T(C_i)}.$$

where $\eta(k_j)$ - number of connections of concept k_j, $T\eta(k_j)$ – total number of connections in cluster C_i.

After determining the weights of the most influential concepts, the LLM can generate a cluster name based on these key terms.

Example of Usage

As a result of analyzing texts using the proposed method, a dataset of documents related to cybersecurity was examined. Based on the extracted quotes, a semantic network was constructed, containing over 50 authors and 200 concepts (Fig. 1).

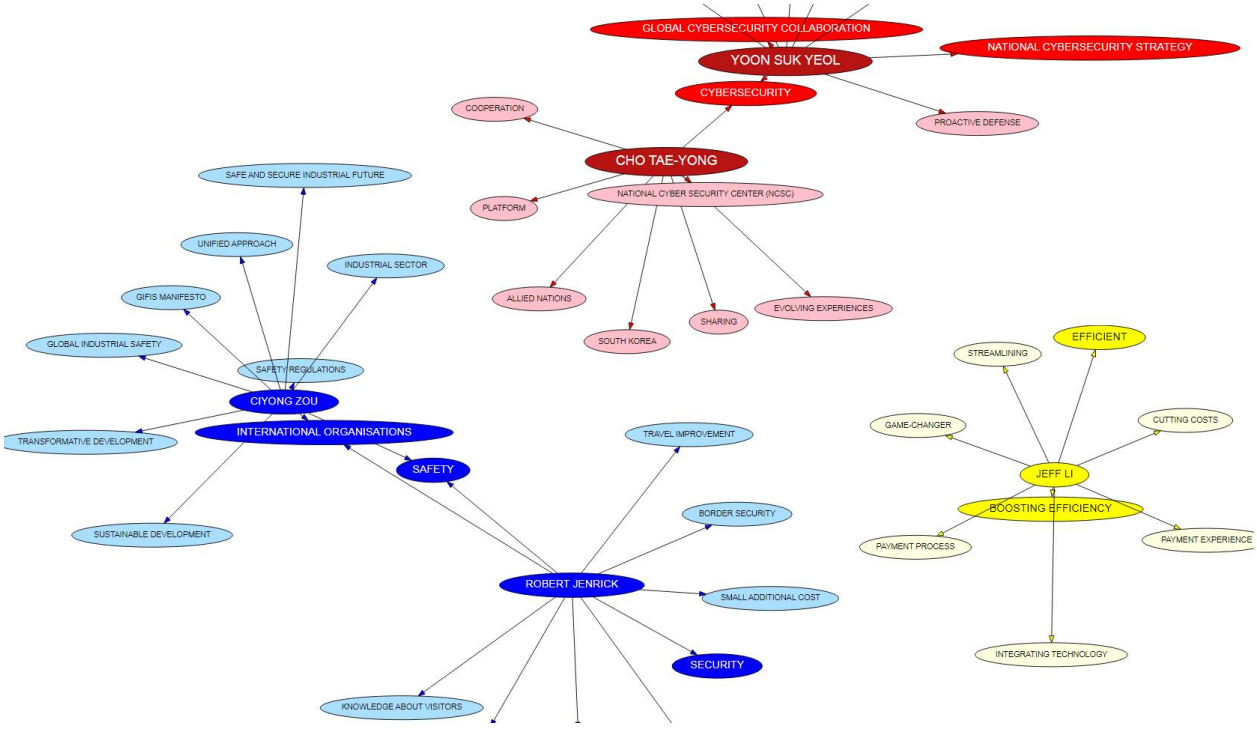

Figure 1. Fragment of the clustered network

Here are examples of authors and quotes:

{
"title": "Yoon reaffirms Seoul's commitment to global cybersecurity cooperation",

```
  "url":
"https://www.koreatimes.co.kr/www2/common/viewpage.asp?newsIdx=382368&categoryCode=205",
  "date": "2024-09-11T10:47:00T",
    "citations": [
      {
        "author": "Yoon Suk Yeol ",
        "citations": [
       "South Korea has long been a stronghold in cybersecurity, developing its defense capabilities and security systems in response to cyberattacks from hostile forces, including North Korea. We will actively contribute to protecting humanity\'s safety and prosperity by sharing our capabilities and experiences with the world",
        "Countries around the world are shifting towards \'active cybersecurity\' based on international solidarity. South Korea also announced its national cybersecurity strategy in February this year, enhancing its proactive defense capabilities to respond preemptively to cyberthreats, and is making concerted efforts to collaborate internationally against transnational cyberthreats",
        " South Korea plans to deepen its cooperation with NATO and reaffirmed his commitment to enhancing cybersecurity collaboration with the alliance"
        ]
      },
      {
        "author": " Cho Tae-yong",
        "citations": [
       "South Korea has been continuously enhancing its cybersecurity capabilities since establishing the National Cyber Security Center (NCSC) in 2004. As cooperation with allied nations is particularly crucial for cybersecurity, we hope that this event serves as a platform for sharing evolving cybersecurity experiences and solutions"
            ]
      },
{
        "author": " Mart Noorma "
        "citations": [
       "South Korea plans to deepen its cooperation with NATO and reaffirmed his commitment to enhancing cybersecurity collaboration with the alliance",
          "Quote 2"
        ]
      }
  ]
}
```

Discussion

This technology opens up new opportunities for automated quote analysis, particularly in large text corpora. The use of a "swarm of virtual experts" to refine concepts enhances the results of the analysis. Automated clustering and cluster naming provide greater context for analyzing thematic groups.

The implementation of direct speech extraction technology can be integrated into a large information and analytical system, with two possible approaches. The first approach involves adding extraction results to each document in a retrospective database, allowing for direct searching within quotes. This global approach requires setting up a dedicated LLM implementation for such a system, offering advantages in flexibility and control over text processing but complicating the use of rapidly evolving LLM models.

The second approach involves using extraction technology and semantic network construction in real-time, only for documents relevant to the user's query. In this case, the retrospective database is not indexed at the quote level, and searching within direct quotes is not possible, but this approach avoids the need to deploy a dedicated LLM. It allows for the use of existing APIs to access modern models, reducing resource costs and increasing the system's implementation efficiency.

Conclusions

In this paper, we examined a method for the automatic extraction and analysis of direct speech from texts using large language models (LLM). The proposed approach enables the extraction of quotes from documents and the creation of networks of authors and concepts based on these direct quotes, opening new opportunities for analysis and research in various fields.

The first step involves using LLM to identify quotes and authors, which automates the extraction of important information from texts. Subsequently, the models perform key concept extraction, helping to understand the main themes of

each quote and create a bimodal network of authors and concepts. This approach allows researchers to uncover connections between different authors based on shared concepts and interests, facilitating a deeper analysis of the text's themes.

Network clustering, carried out by LLM, highlights groups of authors with shared concepts, thereby dividing them into thematic clusters. Using LLM for clustering provides flexibility in choosing methods and allows for the automatic generation of cluster names, reflecting their content and themes.

The proposed technology can be integrated into broader information and analytical systems with two implementation options: either indexing the entire retrospective database for quote searching or operating in real-time for relevant documents without the need for archive indexing. Each of these approaches has its advantages and challenges, depending on the tasks and resources.

Thus, leveraging LLM for automating direct speech extraction and analyzing author concepts significantly speeds up the information processing, provides deep analysis of text data, and creates new opportunities for research and intelligence gathering.

References

[1] Tim O'Keefe . Extracting and Attributing Quotes in Text and Assessing them as Opinions. Doctor of Philosophy thesis, School of Information Technologies Faculty of Engineering & IT The University of Sydney 2014. URL: https://core.ac.uk/download/pdf/41239677.pdf

[2] Andrew Salway, Paul Meurer, Knut Hofland and Øystein Reigem. Quote Extraction and Attribution from Norwegian Newspapers. Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 293–297, Gothenburg, Sweden, 23-24 May 2017. c 2017 Linkoping University Electronic Press. URL: https://aclanthology.org/W17-0241.pdf

[3] Insa Mannstadt, Susan M. Goodman, Mangala Rajan, Sarah R. Young, Fei Wang, Iris Navarro-Millán, Bella Mehta. A Novel Approach for Mixed-Methods Research Using Large Language Models: A Report Using Patients' Perspectives

on Barriers to Arthroplasty. ACR, 2024, Volume 6, Issue 6. doi: 10.1002/acr2.11662.

[4] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, Hong Shen. "I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. arXiv:2403.19876. doi: 10.48550/arXiv.2403.19876

[5] Dmytro Lande, Minglei Fu, Wen Guo, Iryna Balagura, Ivan Gorbov & Hongbo. Yang. Link prediction of scientific collaboration networks based on information retrieval. World Wide Web : Internet and Web Information Systems. - N 23, pp. 2239-2257 (2020). doi: 10.1007/s11280-019-00768-9

[6] Zgurovsky, M., Lande, D., Boldak, A. et al. Linguistic Analysis of Internet Media and Social Network Data in the Problems of Social Transformation Assessment. Cybernetics and Systems Analysis (2021). Volume 57, issue 2. Pages: 228 - 237. doi: 10.1007/s10559-021-00348-8

[7] Nulty, PG. Semantic Network Analysis of Contested Political Concepts. International Conference on Computational Semantics (IWCS 2017). doi: 10.17863/CAM.14415

[8] Alicja Rachwał, Emilia Popławska, Izolda Gorgol, Tomasz Cieplak, Damian Pliszczuk, Łukasz Skowron, and Tomasz Rymarczyk. Determining the Quality of a Dataset in Clustering Terms. Appl. Sci. 2023, 13(5), 2942. doi: 10.3390/app13052942

[9] Lande, D., & Strashnoy, L. (2023). GPT Semantic Networking: A Dream of the Semantic Web–The Time is Now. Kyiv, Engineering Ltd. ISBN - 168 p. ISBN: 978-966-2344-94-3

Appendix

A set of example prompts for each stage of the direct speech extraction and author clustering method you are working with.

1. **Quote Extraction Stage**

At this stage, the LLM should identify authors and their direct quotes within the text. The prompt should include a request to find quotes and authors in the document.

Prompt for extracting quotes:

```
Find all direct quotes in the following text along with the names of the authors who said them. For each quote, provide the author and the quote itself. Format the result as a list of pairs {author: quote}
```

Output:

```
{
 "Author 1": "Quote by Author 1",
 " Author 2": " Quote by Author 2",
 ...
}
```

2. Key Concept Extraction Stage

The LLM is required to identify key concepts (words or phrases) that best describe the content of the quotes.

Prompt for key concept extraction: "Identify key words or concepts for each of the following quotes. Specify the most important words or phrases that best reflect the content of the quote."

Example text for processing:

```
{
 "Author 1": "Quote by Author 1",
 " Author 2": " Quote by Author 2",
}
```

Output:

```
{
 " Author 1": ["Key Concept 1", " Key Concept 2"],
 " Author 2": ["Key Concept 3", " Key Concept 4"]
}
```

3. Stage of Semantic Network Construction

Prompt for building the semantic network:

Create a bimodal network where nodes represent authors and key concepts, and the connections between nodes reflect the relationships between authors and concepts found in their quotes.

Example of input data:

```
{
  "Author 1": ["Key Concept 1", "Key Concept 2"],
  "Author 2": ["Key Concept 3", "Key Concept 4"]
}
```

Output:

```
{
 "nodes": [
  {"id": "Author 1", "type": "author"},
  {"id": "Author  2", "type": "author"},
  {"id": " Key Concept 1", "type": "concept"},
  {"id": "Key Concept 2", "type": "concept"},
  {"id": "Key Concept 3", "type": "concept"},
  {"id": "Key Concept 4", "type": "concept"}
 ],
 "edges": [
  {"from": "Author 1", "to": "Key Concept 1"},
  {"from": "Author 1", "to": " Key Concept 2"},
  {"from": "Author 2", "to": " Key Concept 3"},
  {"from": "Author 2", "to": " Key Concept 4"}
 ]
}
```

4. Clustering Stage

LLM performs clustering based on the constructed semantic network, identifying groups of authors with shared interests through common concepts.

Prompt for Clustering:

```
Cluster the following network of authors and concepts based on their connections. Group authors who share common concepts. Use a clustering method of your choice, such as Louvain or another suitable method
```

Example Input Data:

```
{
 "nodes": [
  {"id": "Author 1", "type": "author"},
  {"id": "Author 2", "type": "author"},
  {"id": "Key Concept 1", "type": "concept"},
  {"id": "Key Concept 2", "type": "concept"},
  {"id": "Key Concept 3", "type": "concept"},
  {"id": "Key Concept 4", "type": "concept"}
 ],
 "edges": [
  {"from": "Author 1", "to": "Key Concept 1"},
  {"from": "Author 1", "to": "Key Concept 2"},
  {"from": "Author 2", "to": "Key Concept 3"},
  {"from": "Author 2", "to": "Key Concept 4"}
 ]
}
```

Output:

```
{
 "clusters": [
```

```
  {
    "cluster_id": 1,
    "authors": ["Author 1", " Author 2"],
    "concepts": ["Key Concept 1", " Key Concept 2"]
  },
  {
    "cluster_id": 2,
    "authors": ["Author 3"],
    "concepts": ["Key Concept 3", " Key Concept 4"]
  }
 ]
}
```

5. Cluster Name Generation Stage

Prompt for generating cluster names:

```
Generate a name for each cluster of authors using their key concepts. The name should reflect the shared theme or interests of the authors in the cluster.
```

Example input:

```
 {
"clusters": [
  {
    "cluster_id": 1,
    "authors": ["Author 1", " Author 2"],
     "concepts": ["Key Concept 1", " Key Concept 2"]
  },
  {
    "cluster_id": 2,
    "authors": ["Author 3"],
```

```
    "concepts": ["Key Concept 3", " Key Concept 4"]
   }
 ]
}
```

Output:

```
{
 "cluster_names": [
   {
    "cluster_id": 1,
    "name": " Common Themes: Concepts 1 and 2"
   },
   {
    "cluster_id": 2,
    "name": " Innovative Approaches: Concepts 3 and 4"
   }
 ]
}
```