Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science for Digital Media

Information Extraction from Academic Mailing Lists using Large Language Models

Master's Thesis

Nazifa Kazimi
Born Sept 12, 1996 in Afghanistan

Matriculation Number 123688

1. Referee: Prof. Dr. Benno Stein
2. Referee: Jun.-Prof. Dr. Jan Ehlers

Submission date: July 31, 2024

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Krefeld, July 31, 2024

.....
Nazifa Kazimi

Abstract

This thesis is centered around the extraction of information from academic mailing lists, with a particular focus on the IR Anthology, a collection of IR-related research papers. The objective is to enhance the IR Anthology by incorporating additional insights and details. While several resources could contribute to this goal, the utilization of archives from mailing lists—a recognized communication channel among scholars—is specifically employed in this research. Recent developments, such as the introduction of OpenAI's language models, have made more advanced methods available for mining and extracting data from texts. Subsequently, the progression of the research involves the utilization of large language models, such as ChatGPT, to extract valuable information from emails for enriching the IR Anthology.

Contents

Chapter 1

Introduction

The Information Retrieval Anthology (IR Anthology) is a comprehensive collection of IR-related publications developed by the Webis research group. As of now, the IR Anthology contains 62,846 papers on the study of information retrieval, serving as a vital repository, gathering communication among IR researchers and spreading research findings within the community.

Academic mailing lists, similarly, act as an essential communication channel where researchers exchange up-to-date news and valuable information. Given the shared goal of these platforms in enhancing scholarly communication, integrating insights from academic mailing lists into the IR Anthology emerges as a logical and beneficial step. To initiate the analysis, the SIGIR mailing list archive is selected for examination. This archive provides a rich source of data, encompassing about 13,000 emails related to the information retrieval community.

In the first exploratory data analysis phase, a preliminary review of the archive is conducted to determine the overall structure and themes of the emails. This preliminary review highlights the diversity of emails, ranging from job announcements to discussions about conferences and journals. These insights lead to the development of a classifier to categorize the emails in a way that would facilitate a more fine-grained analysis of each category.

Categorization of the emails is thoroughly conducted to ensure a comprehensive understanding of the content. Careful attention to detail and significant effort are required for the analysis to explore the details within each category. Classification of these emails is involved in the initial step to gain deeper insights into them. Upon reviewing the emails, two primary labels are considered in the classification process: job-posting and venue-related. The job-posting label is assigned to emails specifically announcing job opportunities, while the venue-related label encompasses emails related to conferences or journals. Emails that do not fall into either category, i.e., neither a job announcement nor associated with a venue, are categorized under the 'other' label. This labeling system provides a structured approach to classifying and distinguishing between different types of emails, facilitating a more systematic analysis. The methodology behind the email classification is detailed in Chapter 3. Building on the structure of the IR Anthology and insights gained from the initial analysis, more fine-grained information is extracted using OpenAI's language models in a subsequent step. This step aims to leverage advanced text processing capabilities to unveil deeper insights into the content of the emails.

Figure 1.1 provides an overview of the systematic approach used in this thesis. The methodology comprises five key stages: Initial Research, Exploratory Data Analysis, Email Classification, Information Extraction, and Evaluation.

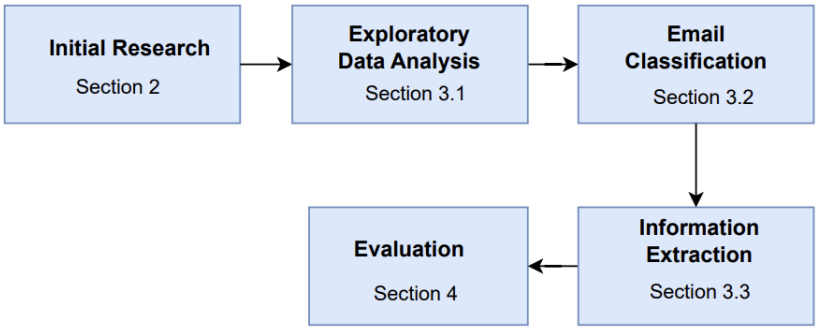

Figure 1.1: Methodology Flowchart

The process begins with Initial Research, which involves a comprehensive literature review of IR-related research, focusing on information extraction techniques and their relevance to the current study.

Next, the Exploratory Data Analysis phase is launched to identify patterns and structures within the data. Email Classification follows, utilizing different techniques such as Keyword-matching, Logistic Regression, and ChatGPT to classify the emails, which is essential for subsequent analysis.

In the Information Extraction phase, the process delves deeper into each email category to determine which details are valuable for inclusion in the IR Anthology. This phase focuses on extracting relevant information using OpenAI's model.

Finally, the Evaluation stage uses a ground truth dataset to assess the performance of the information extraction process, ensuring the reliability of the methodology and leading to robust conclusions. The subsequent chapters provide a detailed explanation of each step.

Chapter 2

Literature Review

Extracting structured information from unstructured text is a major area of natural language processing (NLP) and has been applied in several use cases. Various methodologies are proposed for developing approaches from basic rule-based systems to more recent and more complex deep learning models. This section highlights relevant research in the era of textual information extraction.

The early methods of information extraction heavily relied on rule-based systems for extracting relevant information from texts. However, domain specificity necessitated a lot of manual work in creating functional rules for such systems. The research paper Appelt et al. [1993] is an important milestone in this regard. The study presents FASTUS, an information extraction system based on finite-state machines. FASTUS utilizes a multi-stage pipeline to parse text which includes tokenization, handling of complex words, clause identification as well as phrase recognition and template filling. It enables one to extract structured data from vast amounts of text in an efficient and scalable manner. Compared to full syntactic parsers, text processing has lower computational demands due to the design of the FSM-based architecture. The fact that FASTUS proved that it is effective to split complex activities into smaller sequential steps profoundly affected later information extraction systems. While FASTUS and similar early methods laid the groundwork for text processing, this thesis employs more modern approaches that integrate advanced language models.

The field of information extraction has witnessed significant advancements in recent years, largely driven by the evolution of machine-learning models. These models involve annotated datasets which results in flexible and scalable information extraction. The paper "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" Lafferty et al. [2001] introduces Conditional Random Fields (CRFs), which is a probabilistic graphical model whose main aim is to segment and label sequence data. In the context of Named Entity Recognition (NER), CRFs work more effectively since they consider small groups of words within a sequence and thus model dependencies between labels. By leveraging the contextual dependencies within the data, CRFs improve the accuracy and reliability of extracted information. In information extraction tasks, such as named entity recognition, CRFs can be used to identify entities like names, dates, and locations within a text. There are some popular libraries out there that are used for Named Entity Recognition tasks such as NLTK Bird et al. [2009], Spacy Honnibal and Montani [2017], and Stanford NLP.

The BERT model introduced by Devlin et al. [2018] is pivotal for information extraction from text due to its bidirectional transformer architecture, which captures con-

text from both directions, enhancing the accuracy of entity and relationship extraction. BERT's pre-training on vast corpora, followed by fine-tuning on specific tasks, allows it to achieve state-of-the-art results in various NLP benchmarks, including named entity recognition. Its flexibility and robust performance make BERT adaptable to diverse IE tasks, significantly improving the efficiency and accuracy of extracting structured information from unstructured text.

While early methods such as rule-based systems, FASTUS, CRFs, and advanced models like BERT have made significant contributions to the field of information extraction, this thesis utilizes a different approach. Instead of relying on these traditional or earlier models, the research focuses on leveraging OpenAI's language models.

Large Language Model Techniques involve using advanced machine learning models, known as large language models (LLMs), for information extraction tasks. These models can be used to extract structured data from unstructured text data in a few ways. They are capable of identifying the key entities and relationships in the text. They are able to recognize the names, locations, and events that are discussed in the text, as well as the relations between them. They can also be applied to the unstructured text data in order to extract particular fields of information. They extract, for instance, people's names and event dates. With this data, a structured dataset that is simpler to work with and evaluate can be produced. These models, such as GPT-1 Radford et al. [2018], GPT-2 Radford et al. [2019], GPT-3 Brown et al. [2020], GPT-4 OpenAI [2023] series or Google's BERT Devlin et al. [2018], have been pre-trained on vast amounts of text data and are capable of understanding and generating human-like text. Understanding and extracting complex scientific information from unstructured text poses a significant challenge, especially for individuals lacking experience in natural language processing.

Dunn et al. [2022] present a novel approach for structured information extraction from complex scientific texts using fine-tuned large language models. Their method leverages advanced language models to parse and extract detailed, structured data from scientific literature, which traditionally poses challenges due to its complex and varied formatting. The study demonstrates significant improvements in extraction accuracy and efficiency compared to conventional techniques. By fine-tuning models specifically for scientific texts, the authors achieve better performance in identifying and organizing key information. This work highlights the potential of large language models to enhance data extraction processes in scientific research. However, this thesis does not employ the method presented by Dunn et al., as it focuses on utilizing OpenAI's ChatGPT without fine-tuning for information extraction from academic mailing lists.

Goel et al. [2023] address the challenge of extracting vital patient-related information from unstructured clinical notes in electronic health records. It introduces an innovative approach that combines large language models with human expertise, significantly reducing the time and cost associated with traditional human-centric annotation methods. The study focuses on medical information extraction and demonstrates that this collaborative approach maintains high accuracy levels while minimizing human intervention. It highlights the efficiency of LLMs in accelerating medication information extraction, achieving comparable accuracy to trained medical NLP annotators. Notably, the integration of LLMs into a human-in-the-loop process proves beneficial, generating expert-level annotations and saving considerable human time. The paper looks forward to future work, considering fine-tuning LLMs for specific tasks and the integration of constrained decoding for improved performance. Overall, the study anticipates further advancements in the field, emphasizing the potential of LLMs in enhancing the utility of unstructured clinical

data for the rapid deployment of tailored NLP solutions in healthcare. While Goel et al.'s approach effectively combines LLMs with human expertise, this thesis focuses solely on the use of OpenAI's ChatGPT for information extraction.

Caufield et al. [2024] proposes a method for structured information extraction using a large language model such as GPT-3 involving the following steps: generating a structured prompt given a templated schema, obtaining and completing the prompt with text input, processing the readout response of the model, grounding extracted entities to identifiers in existing semantic web ontologies, and optionally translating the extracted entity grounding results into an ontological representation. The performance of the system is evaluated by grounding entities in the gene ontology and disease ontology structure and is found to be a better grounding method than that of simple prompting. Further evaluation of SPIRES on the Biocreative Chemical-Disease-Relation task demonstrated the zero-shot learning capabilities of falling back to the previously mentioned PICO-like ontology in this case. SPIRES offers a way to systematically extract structured data from unstructured text, and can be adapted for a variety of different use cases. We utilized OpenAI's ChatGPT without integrating ontological schemes, as our simpler approach adequately meets our needs.

Prompt engineering is an important task in the application of OpenAI's models to generate insights from text. Prompt engineering refers to creating prompts that effectively guide the model in producing the desired output. It has been found that the quality and structuring of a prompt are vital to the general performance of a language model in the context of information extraction tasks. Research has recently been dedicated to developing purpose-built skills and strategies for proper prompt engineering. The paper "Making Pre-trained Language Models Better Few-shot Learners" by Gao et al. [2020] presents a novel approach called "AutoPrompt" to enhance the few-shot learning capabilities of pre-trained language models. In this study, AutoPrompt is constructed to enhance the few-shot learning potential of the pre-trained models by optimizing the prompt design. They stress that with the proper design of prompts, there can be some massive improvements and gains across various benchmarks on NLP. This research provides an important look into how one can leverage a pre-trained model to any new task with a small amount of labeled data and bring in a specialized prompt to inform it how to best perform on the newly designed training distribution. The research conducted in prompts from this study is significant because it shows that developing a prompt to inform the pre-trained model on how to have a successful outcome. A catalog of prompt patterns has been discussed in White et al. [2023] that can be used to improve the performance of ChatGPT. Given the importance of prompt engineering, this research focuses on the details of prompts when using OpenAI's models to ensure effective information extraction from academic mailing lists.

In Chapter 3, different prompting methods will be explained, including techniques for both zero-shot and few-shot learning, as well as strategies for optimizing prompts to enhance model performance. Detailed examples and case studies will illustrate the practical application of these methods. Evaluations are included in Chapter 4, where the effectiveness of various prompting techniques will be assessed. Additionally, Chapter 4 will compare the results of different prompting strategies to provide insights into their relative strengths and weaknesses.

Chapter 3

Methodology

In this chapter, we delve into the comprehensive methodology utilized for this research. Our goal is to systematically explore and extract valuable insights from a substantial collection of academic mailing lists. The methodology begins with an Exploratory Data Analysis (EDA) to understand the dataset's structure and content, followed by detailed classification and extraction processes.

3.1 Exploratory Data Analysis

The first and essential phase of working with this archive is conducting an Exploratory Data Analysis (EDA) to locate information that can be useful for enhancing the IR Anthology. After data acquisition of the SIGIR mailing list archive, as we do with BeautifulSoup, a Python library for web scraping, this process begins with a comprehensive examination of the almost 13,000 emails to understand their structure and content, enabling the identification of key themes and categories. To facilitate a systematic analysis, the emails need to be tagged and categorized based on their content. This categorization helps in organizing the emails into distinct groups, making it easier to extract relevant information for the IR Anthology. The EDA on the SIGIR mailing list revealed the following categories of emails:

- **Position Announcements:** Emails announcing job openings, scholarships, internships, and other academic positions.
- **Journal and Conference Announcements:** Emails related to calls for papers, conference details, workshop notifications, and journal announcements.
- **Other Announcements:** Emails announcing awards, books, recognitions, and miscellaneous announcements not directly related to positions or conferences.

3.2 Email Classification

The email classification phase involves categorizing approximately 13,000 emails into distinct classes to facilitate the extraction of relevant information. The primary classes considered in this classification process are "job-posting", "venue-related", and "other". Since the vocabulary used in the first two classes is rather distinct, we expect good classification results and even more simple classification approaches. However, we strive for

an excellent classification result as this is crucial for the subsequent fine-grained information extraction phase. To accomplish the task, several techniques are employed. All these classification techniques are employed on the subjects of emails to categorize and tag the emails accordingly. Initially, a keyword-matching method is used to identify specific terms associated with each class. Following this, a Logistic Regression classifier and OpenAI's ChatGPT are applied to further automate the classification process.

3.2.1 Keyword-Matching Method

The first technique used for email classification is the keyword-matching algorithm, which serves as the primary method for categorizing emails into distinct classes, namely "venue-related", "job-posting", and "other". This algorithm operates on the principle of identifying specific keywords within the email subject to assign each email to one of these predefined categories.

For classifying emails into the "venue-related" category, we try to come up with a set of keywords related to academic and event-related contexts. Conversely, for the "job-posting" category, we compile a list of terms associated with employment and academic positions. We iteratively apply the classifier, manually assess its performance on a random sample of the emails, and modify the keyword list until we could not spot any obvious modification anymore.

The final keywords used for 'venue-related' are: 'cfp', 'submission', 'paper', 'call for participation','participation', 'call-for-participation', 'cpf', 'tutorial', 'tutorials', 'proposal', 'proposals', 'bids', 'workshop', 'conference', 'challenge', 'issue', 'call for posters'.

The final keywords used for job-posting are: 'position', 'job',' assistant', 'scholarship', 'postdoc', 'studentship', 'post-doctoral', 'internship', 'vacancy', 'candidate', 'vacancies', 'opportunity', 'associate', 'post'.

The output of this automated classification process is depicted in Figure 3.1, which illustrates the distribution of emails across the specified categories. This visualization provides an overview of how the keyword-matching algorithm partitioned the email dataset to facilitate further analysis.

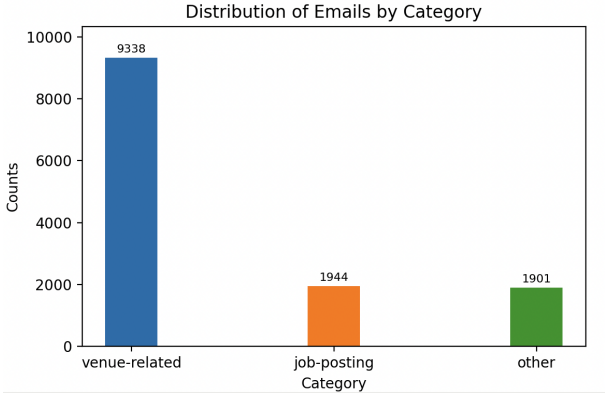

Figure 3.1: Distribution of Emails by Category via Keyword-Matching

In order to assess the classification performance, we sample 100 emails classified as "venue-related" and 100 emails classified as "job-postings". After manual assessment, we find that all 200 emails have been classified correctly. After that, all of the 1,901 emails classified as 'other' are reviewed manually. Here, we find that 1,609 of the emails are incorrectly classified. 1,415 are in fact "venue-related", 194 in fact "job-postings". Hence, overall, the keyword-matching algorithm classifies the emails with an accuracy of $\frac{9338+1944+292}{13183} \approx 85.4\%$. Figure 3.2 shows the distribution of emails by category after manual assessment of the "other" class.

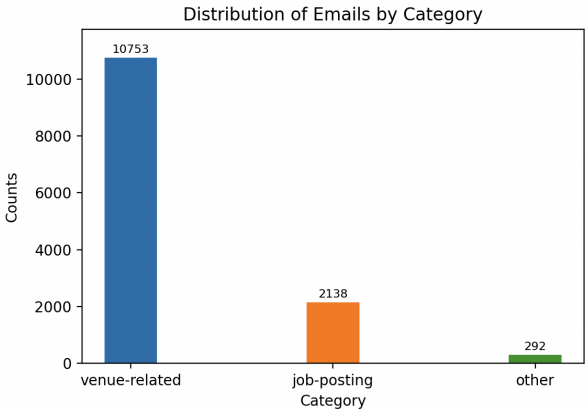

Figure 3.2: Categorization of Emails after Keyword-Matching and Manual Labeling

Note that we use this classification as ground truth for the evaluation of the two following approaches. By establishing this ground truth, future research can build upon a reliable foundation, facilitating a more accurate and meaningful analysis of email categorization techniques.

3.2.2 Logistic Regression

With the establishment of a ground truth dataset, it becomes feasible to train classifiers on a training split of the data and subsequently evaluate it on a hold-out test set. We feed the subjects of emails into the model to perform the classification task. Along these lines, we employ a Logistic Regression classifier, a commonly used method in text classification tasks. Logistic Regression (LR) is primarily known for binary classification, but it can also be adapted for multi-class classification problems. In multi-class classification, LR can handle more than two classes through techniques such as One-vs-Rest (OvR) and Softmax (or Multinomial) Logistic Regression. The OvR approach breaks the problem into multiple binary classifiers, each distinguishing one class from the others. The default One-vs-Rest method is used for the classification of emails into our three classes job-posting, venue-related, and 'other'.

This classification approach involved dividing the dataset, with 80% used for training the model and 20% reserved for testing. On the test set, the trained model demonstrated a high degree of accuracy, achieving a rate of 98%. This convincing result underscores the efficacy of Logistic Regression in automating the categorization process. Consequently, the model can be effectively utilized as a tagging mechanism for future mailing list archives, ensuring consistent and reliable classification of emails.

The application of this machine learning technique not only improves the systematic organization of the diverse information contained within the emails but also facilitates

the scalable and efficient processing of extensive email data across various mailing lists.

3.2.3 ChatGPT-Powered Email Classification

Although we are able to achieve excellent accuracy with Logistic Regression, we are interested in how well ChatGPT, a powerful large language model developed by OpenAI, can be prompted to classify emails by leveraging its advanced natural language processing capabilities. By prompting the model with a set of subject lines, we advise ChatGPT to assign the appropriate label for each email. I.e., the task is to classify the email subjects into three categories: 'job-posting', 'venue-related', and 'other'.

Initial prompt: *"Given a list of email subjects, classify them into one of the three categories: 'job-posting', 'venue-related', or 'other'. If the subject mentions a venue like a conference or a journal, label it as 'venue-related'. Subjects related to announcements not tied to conferences or journals, such as releases or book announcements, are labeled 'other'. Each subject falls into only one category, and there are no titles included in the output. Return labels in CSV format. "*

Challenges and adjustments of initial prompt

The initial prompt provided general instructions for classifying email subjects into three categories. First, the model has been given the initial prompt.

Sample result of initial prompt:
1. [SIG-IRList] Okapi released under BSD License

- Ground Truth Label: other
- Predicted Label: `job_posting`

2. [SIG-IRList] ANN: S-Match Open Source Updated

- Ground Truth Label: other
- Predicted Label: `venue_related`

3. [SIG-IRList] Student grants deadline extended - Autumn School 2016 for Information Retrieval and Information Foraging (ASIRF) in Germany

- Ground Truth Label: other
- Predicted Label: `job_posting`

4. [SIG-IRList] [IAAIL] [ICAIL2019] LegalAIIA 2019: Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace – 2nd Call for Papers (CfP)

- Ground Truth Label: `venue_related`
- Predicted Label: `job_posting`

The outcome, around 66.00% with the initial prompt, suggests that the prompt needs to be changed. In light of the information given here, the prompt ought to include more information and procedure phases. Therefore, to achieve a better outcome, the prompt should have more specifics and distinct phases, according to prompt-engineering methodologies.

Revised prompt: *Classify email subjects into 'job-posting', 'venue-related', or 'other'. Consider the following criteria: 'other': Subjects related to lectures, surveys, awards, dataset releases, software release, and books should be labeled 'other'. Look for keywords such as lecture, survey, award, dataset, book, publication, release, author, and exclude subjects with the '[SIG-IRList]' prefix. 'venue-related': Subjects mentioning calls for papers (cfp), conferences, workshops, journals, or community events are 'venue-related'. Look for keywords like cfp, conference, workshop, journal, and specific criteria. Exclude subjects with the '[SIG-IRList]' prefix. 'job-posting': Subjects indicating job opportunities or positions are 'job-posting'. Look for keywords such as job, position, PhD, Postdoc, studentship. Exclude subjects with the '[SIG-IRList]' prefix. Return only labels in CSV format.*

The revised prompt offers clearer, more comprehensive guidance for grouping email subjects into designated groups. The modifications are intended to raise the model's comprehension and raise the classification's accuracy. The main changes and their objectives are outlined below:

- 'other': It includes subjects related to lectures, surveys, awards, dataset releases, software releases, and books. The adjusted prompt instructs us to look for keywords such as lecture, survey, award, dataset, book, publication, release, and author. Additionally, it emphasizes excluding subjects with the '[SIG-IRList]' prefix.
- 'venue-related': The adjusted prompt specifies that subjects mentioning calls for papers (cfp), conferences, workshops, journals, or community events should be labeled 'venue_related'. It provides keywords like CFP, conference, workshop, journal, and specific criteria. Similar to 'other', it instructs to exclude subjects with the '[SIG-IRList]' prefix.
- 'job-posting': The adjusted prompt directs to identify subjects indicating job opportunities or positions. It provides keywords such as job, position, PhD, Postdoc, and studentship. It also emphasizes excluding subjects with the '[SIG-IRList]' prefix.

Overall, the adjusted prompt provides more detailed instructions and additional keywords for better categorization. The exclusion of [SIG-IRList] is because its mention does not indicate any venue. With this adjustment, the accuracy of tagging is around 90.00%.

The lower performance of ChatGPT compared to the more traditional Logistic Regression model can be attributed to several factors. ChatGPT is a general-purpose model and may not be as finely tuned to the specific nuances of the email classification task. Additionally, the performance of ChatGPT heavily depends on the quality and specificity of the prompts used, which can lead to variability in its accuracy. Logistic Regression, being a simpler and more specialized model, can achieve higher accuracy in this particular application due to its prior training on this task.

3.3 Information Extraction

Given that the email dataset is now divided into three classes, we are interested in whether more fine-grained information can be extracted from the emails in the classes "venue-related" and "job postings", which can then be used to augment the IR Anthology. The objective of this section is to describe the methodology used to extract information from the labeled emails.

3.3.1 Data Preparation

This subsection details the methodology used to prepare the dataset. With the dataset now classified into two primary categories, namely job-posting and venue-related emails, the focus of the analysis shifts towards these two distinct groups. While the job postings subset is kept as is, for the venue-related emails, the goal is to conduct a detailed examination while ensuring the analysis remains relevant and manageable. To achieve this, the venue-related emails are further filtered to include only those that reference specific conferences indexed in the IR Anthology.

The rationale for this filtering process is to concentrate on emails that pertain to well-known and reputable conferences within the IR community. The narrowed dataset includes the conferences ['ADCS', 'AIRS', 'CCIR', 'CERI', 'CHIIR', 'HCIR', 'CIKM', 'CIVR', 'CLEF', 'CORIA', 'DESIRES', 'DIR', 'ECIR', 'FDIA', 'FIRE', 'ICTIR', 'IRFC', 'ISMIR', 'ICMR', 'NTCIR', 'RIAO', 'OAIR', 'SIGIR', 'SPIRE', 'TREC', 'WSDM', 'WWW']. This targeted approach enhances the relevance of the findings and makes the dataset more manageable for detailed analysis.

A subset of these emails is randomly selected and manually annotated to create a ground truth for comparison. The dataset consists of 100 emails, evenly divided between job postings and venue-related emails:

- **Job Postings**: 50 emails
- **Venue-Related Emails**: 50 emails

Each email in the dataset is further annotated manually for specific details. The annotated details for job-posting emails are: *Contact Person, Position, Organization, Group,* and *Work Area*. The annotated details for venue-related emails are: *Topics of Interest,* and *Conference Year*.

To ensure a robust evaluation, the dataset is divided into development (dev) and test sets. The split is done randomly while maintaining an even distribution of email types across both sets:

- **Development Set**: 60 emails (30 job postings and 30 venue-related emails)
- **Test Set**: 40 emails (20 job postings and 20 venue-related emails)

The development set is used to fine-tune the prompts and methods, while the test set is reserved for final evaluation to measure the performance of the models on unseen data.

As mentioned, each email is manually reviewed, and the required details are extracted and recorded in a structured format. This manual annotation served as the ground truth for evaluating the model's performance.

3.3.2 Prompt Construction

Prompt engineering is crucial when working with large language models such as ChatGPT. It is an iterative process to get the desired output by prompt modification. Different techniques can enhance the performance and accuracy of an AI model when extracting information from text. Two common techniques include zero-shot prompting and few-shot prompting. The *Zero-Shot Prompting* involves asking the AI to perform a task without providing any examples. Such a prompt directly instructs the model to perform a task without any additional examples to steer it. The task is simply described to the AI, relying on the model's pre-existing knowledge and understanding. Large language models have been reported to achieve impressive performance in some zero-shot scenarios, but they struggle with other, often complex tasks. To improve their performance, *Few-Shot Prompting* can be used. This technique involves providing a few examples in the prompt, which helps the model learn from the context and perform better. These examples guide the model in generating the desired response.

3.3.3 Performance Evaluation

The performance of the information extraction system is evaluated using four key metrics: Precision, Recall, F1 Score, and Jaccard similarity. These metrics are widely used in information retrieval and natural language processing to measure the accuracy and completeness of extracted information.

- **Precision**: The ratio of correctly extracted information to the total amount of information extracted. It is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.1}$$

- **Recall**: The ratio of correctly extracted information to the total amount of relevant information present in the ground truth. It is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.2}$$

- **F1 Score**: The harmonic mean of Precision and Recall, providing a single metric that balances both. It is defined as:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

- **Jaccard Similarity**: The Jaccard Similarity Index is a reliable technique for calculating the similarity between sets and, consequently, between strings. Because of its ease of use, readability, and efficiency in determining the level of similarity between two sets, this metric has gained a lot of traction. To compare the similarity and diversity of sample sets, statisticians utilize the Jaccard Similarity Index, sometimes referred to as the Jaccard Coefficient. The Jaccard Similarity Index for two sets, A and B, can be calculated by dividing the size of the intersection by the size of the union of the sets. In mathematical notation, it is expressed as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

In the context of string comparison, each string is first tokenized into a set of words or characters (we use word tokens in our evaluations). The Jaccard Similarity Index is then computed based on these token sets.

For example, consider two strings:

- String 1: `"apple orange"`
- String 2: `"apple banana"`

Tokenizing these strings into sets of words, we get:

- Set 1: {`apple, orange`}
- Set 2: {`apple, banana`}

The intersection of these sets is {`apple`}, and the union is {`apple, orange, banana`}. Therefore, the Jaccard Similarity Index is:

$$J = \frac{1}{3} \approx 0.333$$

In the following chapter, we report the performance values of these metrics for the described information extraction task.

Chapter 4

Evaluation

This section focuses on assessing and analyzing the performance of large language models for the task of extracting fine-grained information from venue-related and job postings emails. As a large language model, GPT-3.5-turbo is used, an advanced language model known for its high accuracy and efficiency in natural language processing tasks. With a maximum token limit of 400 per request, GPT-3.5-turbo is prompted to extract specific information from the emails. Note that we evaluate only on the subset of the data which is developed in Section 3.3.1.

4.1 Information Extraction from Venue-Related Emails

For the extraction of information from venue-related emails, the focus is on two key aspects: extracting topics of interest and identifying the year of the conference mentioned in the email content.

The first prompt designed for this purpose is:

P1_1: "Your task is to extract specific details from the provided email.
1. Task descriptions or topics of interest.
2. Year of the conference.
Output Format: Please return the extracted information in the following dictionary format:

```
{'topics_of_interest_pr': ['item1', 'item2'], 'year_pr': 'year'}
```

Special Instructions: Ensure that the extracted information matches the wording in the email without any modifications. The tasks or topics are typically listed as separate items or in a bullet-point format. If any of the required details are not mentioned in the email, return an empty string for that key in the dictionary. Maintain the exact dictionary key structure as shown in the template."

On the development set, prompt **P1_1** achieves an F1 score of 1.0 for the conference year extraction and an F1 score of 0.673 for the extraction of the topics of interest. The Jaccard similarity for the latter is 0.653. To give an intuition of the false predictions the model makes for prompt **P1_1**, a few emails for which the topics of interest are not correctly extracted are listed below. The complete table is in Appendix A, Table A.1.

No	Topics of Interest	Topics of Interest(pr)
1	information (seeking and searching) behaviour (ib), human computer interaction (hci), information retrieval (ir)	user-centered approaches to the design and evaluation of systems for information access, retrieval, and use, information (seeking and searching) behaviour (ib), human computer interaction (hci), information retrieval (ir)
2	innovative and risky information access and retrieval system ideas, systems-building experience and insight, resourceful experimental studies, provocative position statements, new application domains	papers, prototypes, abstracts, open problems in ir
3		tutorials, keynotes, panel session on evaluation initiatives, mirrors special session, regular paper sessions, concerts, social program
4	web search and data mining	web search, data mining

Table 4.1: Emails with incorrectly predicted topics of interest-P1_1

Analysis of Incorrect Predictions of Topics of Interest by P1_1

1. **Misalignment with Provided Topics:** The predicted topics of interest contain more details than actual topics of interest. For example, email *1*.
2. **Completely Wrong Predictions:** The model occasionally predicts completely wrong topics of interest. For example, email *2*.
3. **Empty Actual Topics:** In some cases, the actual topics column is empty, indicating no topics were listed in the email, but the model still predicts topics. For example, email *3*.

Based on the analysis above, it is evident that the model P1_1 exhibits several types of errors in predicting topics of interest. In an attempt to address these errors, we refine the prompt in the following way:

P1_2 = "Your task is to extract specific details from the provided email.
1. Task descriptions or topics of interest.
2. Year of the conference.
Output Format: Please return the extracted information in the following dictionary format:

```
{'topics_of_interest_pr': ['item1', 'item2'], 'year_pr': 'year'}
```

Special Instructions: Ensure that the extracted information matches the wording in the email, without any modifications. The tasks or topics are typically listed as separate items or in a bullet-point format. If any of the required details are not mentioned in the email, return an empty array. Maintain the exact dictionary key structure as shown in

the template. Contextual Clues: Look for phrases such as 'topics of interest include', 'we encourage papers on', 'in the area of', 'papers about', or similar wording to identify relevant topics. Special Case Handling: If the actual topics of interest include items separated by 'and' (e.g., 'search and retail'), do not split them into separate items in the prediction. "

On the development set, prompt **P1_2** achieves an improved F1 score of 0.74, and a Jaccard similarity of 0.721. Again, the table below provides some examples where the topics of interest were still incorrectly predicted. The complete table is in Appendix A, Table A.2.

No	Topics of Interest	Topics of Interest(pr)
1	['total recall problem', 'diagnostic test accuracy (dta) reviews']	['diagnostic test accuracy (dta) reviews']
2	['future directions of information access', 'early research such as pilot studies, presenting challenges and future opportunities, conceptual and theoretical work, and contributions from doctoral work']	['the early research such as pilot studies', 'presenting challenges and future opportunities', 'conceptual and theoretical work', 'contributions from doctoral work']
3	['new or improved models of relevance', 'ranking', 'representation', 'information needs', 'evaluation', 'formal frameworks', 'new search paradigms', 'methods from related disciplines']	['new or improved models of relevance', 'ranking', 'representation', 'information needs', 'evaluation', 'define new tasks', 'develop new search paradigms', 'apply methods from related disciplines']

Table 4.2: Emails with incorrectly predicted topics of interest-P1_2

Analysis of Incorrect Predictions of Topics of Interest by P1_2

1. **Partial Matches:** In some cases, the predicted topics are a subset of the actual topics. For example, emails *1, 2*. In email 1, the model predicted 'diagnostic test accuracy (dta) reviews' but missed 'total recall problem'.

2. **Misinterpretation or Expansion:** The model might misinterpret or expand the scope of the actual topics. For example, in email *3*, the model's prediction includes 'define new tasks' and 'develop new search paradigms' which are not explicitly mentioned in the actual topics but are inferred.

Summary: The enhancements introduced in Prompt P1_2 have resulted in a notable improvement in the model's performance in extracting topics of interest from emails in the development set. The increased precision, recall, and Jaccard similarity demonstrate that P1_2 is more effective in accurately identifying and matching the topics of interest. The specific improvements in handling detailed matches, reducing completely wrong predictions, and effectively generalizing detailed topics have contributed to the overall better performance of P1_2.

Future improvements could focus on further refining the model's ability to handle complex topics and reducing any remaining differences between the actual and predicted

topics. Additionally, continuous refinement of the contextual clues and special case-handling instructions can help enhance the model's accuracy even further.

Table 4.3: Model Performance Evaluation of Venue-Related Emails

Category	Prompt	Data	Precision	Recall	F1 score	Jaccard
year	P1_1	Dev	1.000	1.000	1.000	1.000
topics of interest	P1_1	Dev	0.666	0.681	0.673	0.653
year	P1_2	Dev	1.000	1.000	1.000	1.000
topics of interest	**P1_2**	**Dev**	**0.748**	**0.731**	**0.740**	**0.721**
year	P1_1	Test	1.000	1.000	1.000	1.000
topics of interest	P1_1	Test	0.662	0.664	0.663	0.634
year	P1_2	Test	1.000	1.000	1.000	1.000
topics of interest	P1_2	Test	0.634	0.617	0.625	0.612

As can be observed in Table 4.3, which lists all results for the venue-related emails, prompt P1_2 performs better on the development set, but worse on the test data. This suggests that prompt P1_2 is *overfitting*. Overfitting happens when a model (or in this case, a prompt) captures the noise or specific patterns of the development data too well, leading to a poorer generalization of unseen test data. Prompt P1_1 performs better on the test data, which suggests that it generalizes better to unseen data compared to Prompt P1_2. Generalization is crucial for real-world applications because it indicates how well the model (or prompt) will perform on new, unseen data.

4.2 Information Extraction from Job Announcement Emails

For the extraction of information from job position emails, the focus is on five key aspects: the organization offering the position, any group or department of the organization to which the position is assigned, the contact person, the position title, and the work area.

As with venue-related emails, the prompt used to extract the five aspects is refined in an iterative manner after an error analysis. The three evaluated prompts are given below. Note that the third one uses few-shot prompting, i.e. add examples to the prompt:

P2_1 = "Your task is to extract specific details from the provided email, ensuring that each piece of information is captured exactly as it appears in the email. The details to extract are:
Contact Person: Identify the individual responsible for inquiries, if mentioned.
Position Title: Determine the name of the job or role being offered.
Organization: Extract the name of the organization offering the position.
Group or Department: Specify the group or department related to the position, if mentioned.
Work Area: Identify the primary work area or field associated with the position.
Output Format: Please return the extracted information in the following dictionary format:

```
{
'position_pr': 'position_name',
'organization_pr': 'organization_name',
'contact_person_pr': 'name_of_contact_person',
'group_pr': 'group_or_department_name',
'work_area_pr': 'work_area_name'
}
```

Special Instructions: Ensure that the extracted information matches the wording in the email without any modifications. If any of the required details are not mentioned in the email, return an empty string for that key in the dictionary. Maintain the exact dictionary key structure as shown in the template."

P2_2 = "Your task is to extract specific details from the provided email, ensuring that each piece of information is captured exactly as it appears in the email. The details to extract are:
1. Contact Person: Extract the full name and title of the contact person(s) as provided in the main content of the email not in subsections, avoiding email addresses and ensuring no parts of the name or title are omitted.
2. Position Title: Extract the exact job title from the email content. Ensure that the job title matches the detail level provided exactly and does not include additional information such as department names, work areas, or qualifiers. The title should be verbatim from the email without paraphrasing.
3. Organization: Extract the full name of the organization from the email content, including subunits and departments. "Ensure the name is complete, maintains consistent naming conventions, and avoids abbreviations or acronyms. The organization name

should match exactly how it is provided in the email, without any paraphrasing or omissions.
4. Group or Department: Extract the full and precise name of the specific group or department mentioned in the email. Utilize contextual clues such as units, research groups, or departments to ensure the extracted name is accurate and complete. The name should match exactly how it is provided in the email, without any paraphrasing or omissions.
5. Work Area: Extract the specific work area or work areas mentioned in the email content. Include each area separately if multiple areas are provided, and avoid using broad terms or general organizational focuses unless no specific areas are mentioned. Ensure the extracted work areas match the level of detail provided in the email, without paraphrasing or omissions. Output Format: Please return the extracted information in the following dictionary format:

```
        {
    'position_pr': 'position_name',
    'organization_pr': 'organization_name',
    'contact_person_pr': 'name_of_contact_person',
    'group_pr': 'group_or_department_name',
    'work_area_pr': 'work_area_name'
}
```

Special Instructions: Ensure that the extracted information matches the wording in the email without any modifications. If any of the required details are not mentioned in the email, return an empty string for that key in the dictionary. Maintain the exact dictionary key structure as shown in the template."

P2_3 = ("Your task is to extract specific details from the provided email, ensuring that each piece of information is captured exactly as it appears in the email. The details to extract are:
1. Contact Person: Extract the full name of the contact person(s) mentioned in the content of the email. Ensure the names and titles are complete and accurate, excluding email addresses or contact details. If multiple contacts are mentioned, return all. Do not derive the contact person's name from an email address.
2. Position Title: Predict the position title using the same level of detail as it is provided in the email, without including the work area or unnecessary specifics.
3. Organization: Extract the organization's full name from the email content, including subunits and departments. Ensure the name is complete, maintain consistent naming conventions, and avoid abbreviations or acronyms. The organization name should match exactly how it is provided in the email, without any paraphrasing or omissions.
4. Group or Department: Extract the full and precise name of the specific group or department mentioned in the email. Utilize contextual clues such as units, research groups, or departments for identifying groups. The name should match exactly how it is provided in the email, without any paraphrasing or omissions.
5. Work Area: Extract the specific work area or work areas mentioned in the email content. Include each area separately if multiple areas are provided, and avoid using broad terms or general organizational focuses unless no specific areas are mentioned. Ensure the extracted work areas match the level of detail provided in the email, without paraphrasing or omissions.

Special Instructions: If any of the required details are not mentioned in the email, return an empty string for that key in the dictionary. Maintain the exact dictionary key structure as shown in the example. Consider following examples:
Input1: JUNIOR PROFESSOR IN NATURAL LANGUAGE PROCESSING AND MULTIMEDIA INTERACTION. In the Science, Engineering and Technology Group of KU Leuven (Belgium), Faculty of Engineering Science, Department of Computer Science, there is a full-time tenure-track academic vacancy in the area of natural language processing and multimedia interaction. We seek applications from internationally oriented candidates with an outstanding research track record and excellent didactic skills. The successful candidate will perform research in the Human-Computer Interaction research unit. He or she holds a PhD in Computer Science (or a relevant equivalent degree) with focus on natural language processing and multimedia interaction, and has excellent knowledge of the fundamental principles, algorithms and methods of machine learning. The tenure track of a junior professor lasts 5 years. After this period and subject to a positive evaluation of the tenure track, he or she will be permanently appointed as an associate professor. Postdoctoral researchers are encouraged to apply. More info on the vacancy and instructions on how to apply see: https://www.kuleuven.be/personeel/jobsite/jobs/60022759. You can apply for this professorship till October 15, 2021."
Output1:

```
{
    'position_pr': 'JUNIOR PROFESSOR',
    'organization_pr': 'KU Leuven',
    'contact_person_pr': '',
    'group_pr': 'Science, Engineering and Technology Group',
    'work_area_pr': 'NATURAL LANGUAGE PROCESSING AND MULTIMEDIA INTERACTION'
}
```

Input2: The University of Texas at Austin's School of Information (iSchool) is seeking talented students to join our Ph.D. Program in Fall 2021! Applications are due December 1st, 2020. We seek to engage the best and brightest people who thrive on challenges to enter our Ph.D. program and help shape the future of research. Our research mission is to harness the massive scale and complexity of information, discover the principles and processes to manage it, and leverage information to enhance human lives. We seek to design information management solutions that are accessible, useful, usable, and sustainable. To increase our scientific understanding of the role and impact of information in all human endeavors, we study problems and develop solutions for better information design, management, organization, preservation, and retrieval. Information Retrieval Faculty: Jacek Gwizdka, Matt Lease, Soo Young Rieh, and Yan Zhang For any questions about the doctoral program, please contact our Director of Doctoral Studies, Dr. Yan Zhang <[log in to unmask]>. – Matt Lease Associate Professor School of Information University of Texas at Austin Voice: (512) 471-9350 · Fax: (512) 471-3971 · Office: UTA 5.536 http://www.ischool.utexas.edu/ ml
Output2:

```
{
    'position_pr': 'Ph.D. Program',
    'organization_pr': 'University of Texas at Austin',
    'contact_person_pr': 'Dr. Yan Zhang,Ph.D',
```

```
    'group_pr': 'School of Information (iSchool)',
    'work_area_pr': 'information studies'
}

"
```

In the following, a detailed analysis of the extraction performance across the three prompts is provided for each of the five targeted aspects.

4.2.1 Extraction of the Contact Person

The table below shows a few examples where contact persons are not correctly predicted with prompt P2_1. The complete table is in Appendix A, A.3.

No	Contact Person	Contact Person(pr)
1		unspecified
2		iadh.ounis@glasgow.ac.uk
3	professor ian ruthven (head of sisrg) and/or dr martin halvey (director postgraduate teaching)	dr martin halvey
4	associate professor isabelle augenstein	isabelle augenstein

Table 4.4: Emails with incorrectly extracted contact persons-P21

Error Analysis:

1. **Unspecified Predictions:** There are instances where the predicted contact person is marked as "Unspecified," indicating that the model failed to extract any contact person information from the email. However, the model should only return an empty string.

2. **Email Address Instead of Name:** Occasionally, the model predicts an email address instead of the contact person's name. For instance, the actual contact person is missing in email *2*, and the model predicts the email address (e.g., "iadh.ounis@glasgow.ac.uk") instead of the name.

3. **Ambiguous Listings:** When multiple contact persons are listed, the model sometimes selects only one or fails to capture the full context. For example, email *3*.

4. **Partial Name Extraction:** The predicted contact persons often include only a part of the name. For instance, email *4*.

Performance Evaluation:

- Prompt P2_1:

  1. Contact Person: Identify the individual responsible for inquiries, if mentioned.

  - **Jaccard Similarity:** 0.802
  - **Strengths:** The prompt is straightforward and concise, instructing the model to identify the contact person without any additional constraints.
  - **Weaknesses:** The simplicity might leave some ambiguity regarding what exactly should be extracted (e.g., titles, multiple contacts).

- Prompt P2_2:

  > 1. Contact Person: Extract the full name and title of the contact person(s) as provided in the main content of email not in subsections, avoiding email addresses and ensuring no parts of the name or title are omitted.

  - **Jaccard Similarity:** 0.663
  - **Strengths:** This prompt is more specific, guiding the model to avoid subsections and email addresses while ensuring full names and titles are included.
  - **Weaknesses:** The additional specificity might have inadvertently confused the model, leading to a drop in performance. The mention of "main content of email not in subsections" could be interpreted in various ways, adding complexity.

- Prompt P2_3:

  > 1. Contact Person: Extract the full name of contact person(s) mentioned in the content of the email. Ensure the names and titles are complete and accurate, excluding email addresses or contact details. If multiple contacts are mentioned, return all. Do not derive the contact person's name from an email address.

  - **Few-shot prompting:** Provided two examples to guide the model.
  - **Jaccard Similarity:** 0.838
  - **Strengths:** The few-shot examples likely helped the model understand the prompt better and provided context for what the expected output should look like. The instructions are clear and comprehensive.
  - **Weaknesses:** Although the performance improved, it might still struggle with the complexity if not guided properly by the examples.

4.2.2 Extraction of the Position Title

The table below shows a few examples where the position titles are not correctly predicted with prompt P2_1. The complete table is in Appendix A, A.7.

No	Position	Position(pr)
1	PhD position	PhD positions in Computer Science
2	Tenure-track faculty	Tenure-track faculty in Digital Archives, Preservation, and/or Curation (open rank)

Continued on next page

Table 4.5 – continued from previous page

No	Position	Position(pr)
3	PhD	PhD position: Methods for Review Processes Using AI and ML

Table 4.5: Emails with incorrectly predicted position -P2_1

Error Analysis:

1. **Inclusion of Work Area:** The predicted positions often include the work area, such as specific fields or disciplines (e.g., "PhD positions in Computer Science"), which are already provided in a separate *work_ area* column.

2. **Descriptive Titles:** The predictions often include a detailed description of the role. For instance, "Tenure-Track Faculty" is predicted as "Tenure-Track Faculty in Digital Archives, Preservation, and/or Curation (Open Rank)," which introduces more specifics than the original title.

3. **Contextual Specificity:** The predicted positions sometimes provide more context and specifics than required. For example, a simple "PhD" position is predicted as "PhD Position: Methods for Review Processes using AI and ML," adding unnecessary details.

Performance Evaluation:

- Prompt P2_1:

  2. Position Title: Determine the name of the job or role being offered.

  - **Jaccard Similarity:** 0.391
  - **Strengths:** The prompt is simple and direct, focusing on identifying the name of the job or role.
  - **Weaknesses:** The vagueness of the prompt might lead to varying interpretations of what constitutes the "name" of the job or role. This lack of specificity can result in inconsistent or inaccurate extractions.

- Prompt P2_2:

  2. Position Title: Extract the exact job title from the email content. Ensure that the job title matches the detail level provided exactly and does not include additional information such as department names, work areas, or qualifiers. The title should be verbatim from the email without paraphrasing.

  - **Jaccard Similarity:** 0.289
  - **Strengths:** This prompt is very specific, aiming to capture the job title exactly as it appears without additional details or paraphrasing.

  - **Weaknesses:** The high level of specificity might limit the model's ability to generalize and extract the correct information. The instructions to exclude additional details could lead to missing relevant context or nuances in the job title.

- Prompt P2_3:

  > 2. Position Title: Predict the position title using the same level of detail as it is provided in the email, without including the work area or unnecessary specifics.

  - **Jaccard Similarity:** 0.559
  - **Strengths:** This prompt strikes a balance between being specific enough to avoid unnecessary details and general enough to capture the essence of the job title. Additionally, including examples helps guide the model, providing context and reducing ambiguity in predictions.
  - **Weaknesses:** By providing examples, there is a risk of the model overfitting to those specific examples, which might limit its ability to generalize to new, unseen data.

4.2.3 Extraction of the Work Area

The table below shows a few examples where the work areas are not correctly predicted with prompt P2_1. The complete table is in Appendix A, A.9.

No	Work Area	Work Area (pr)
1	Machine learning, data science, explainable AI (XAI), responsible AI or trustworthy AI	Machine learning / data science
2	Privacy-preserving natural language processing, security and privacy in artificial intelligence (SENPAI)	Privacy-preserving natural language processing
3	Digital archives, digital preservation, digital curation, born digital preservation and data curation, cultural heritage and collections as data, critical archival and data studies, software and platform studies, community-centered digital practices and preservation, critical approaches to metadata, digital accessibility, digital sustainability, digital preservation infrastructures, tools, and policies	Information sciences

Table 4.6: Emails with incorrectly predicted work area -P2_1

Error Analysis:

1. **Single Work Area Limitation**: The prompt requested only one work area, leading the model to choose either one broad field or the most prominent single area. This was evident in complex emails with multiple areas, such as emails *1, 2*.
2. **Focus on Organizational Focus Instead of Work Area**: The model returned the organizational focus instead of the specific work area, such as email *3*.

Performance Evaluation:

- Prompt P2_1:

  > 3. Work Area: Predict the work area or work areas mentioned in the email content. Include each area separately if multiple are provided, and avoid broad terms or organizational focuses unless no specific areas are mentioned.

  - **Jaccard Similarity:** 0.342
  - **Strengths:** This prompt is simple and straightforward.
  - **Weaknesses:** The prompt's simplicity may lead to insufficient specificity. It may lead to incorrect predictions if multiple work areas are mentioned.

- Prompt P2_2:

  > 3. Work Area: Extract the specific work area or work areas mentioned in the email content. Include each area separately if multiple areas are provided, and avoid using broad terms or general organizational focuses unless no specific areas are mentioned. Ensure the extracted work areas match the level of detail provided in the email, without paraphrasing or omissions.

  - **Jaccard Similarity:** 0.372
  - **Strengths:** This prompt points out for multiple work areas.
  - **Weaknesses:** Despite its detailed instructions, the prompt might still struggle with ambiguities related to broad terms and the level of detail. The added requirement to match detail levels exactly could make it more difficult for the model to perform well, leading to a slightly improved but still low Jaccard similarity.

- Prompt P2_3:

  > 3. Work Area: Extract the specific work area or work areas mentioned in the email content. Include each area separately if multiple areas are provided, and avoid using broad terms or general organizational focuses unless no specific areas are mentioned. Ensure the extracted work areas match the level of detail provided in the email, without paraphrasing or omissions.

  - **Few-shot prompting:** Provided examples to guide the model.

- **Jaccard Similarity:** 0.404
- **Strengths:** This prompt, while similar to P2_2, benefits from few-shot examples that help clarify what is expected in terms of detail and specificity. The inclusion of examples likely improved the model's understanding and performance.
- **Weaknesses:** Despite the improvement, challenges remain in handling broad terms and ensuring the exact match of details. The model's performance might still be limited by ambiguities in the prompt's instructions.

4.2.4 Extraction of the Organization Name

The table below shows a few examples where the organization names are not correctly predicted with prompt P2_1. The complete table is in Appendix A, A.12.

No	Organization	Organization (pr)
1	University of Illinois at Urbana-Champaign	University of Illinois, Urbana-Champaign
2	CIRAD (French Agricultural Research Centre for International Development)	CIRAD
3	University of Illinois at Urbana-Champaign	University of Illinois
4	IIT-CNR, Pisa, Italy	The Ubiquitous Internet Research Unit @ IIT-CNR, Pisa, Italy
5	University of Texas at Austin	The University of Texas at Austin's School of Information (iSchool)
6	University of Oklahoma	School of Library and Information Studies (SLIS) at the University of Oklahoma
7	Indiana University Bloomington	Indiana University

Table 4.7: Emails with incorrectly predicted organization -P2_1

Error Analysis:

1. **Partial or Incomplete Names**: The model sometimes returned partial or simplified versions of the organization's name instead of the complete name provided. For example, emails *1, 3*.
2. **Simplification of Organizational Names**: In some cases, the model overly simplified the organization's name, losing critical details. For example, email *2*.
3. **Misinterpretation of Organizational Units**: The model sometimes predicted a specific unit or department within an organization instead of the organization itself. For example, emails *4, 5, 6*.
4. **Omission of Location or Specific Attributes**: The model occasionally omitted location or specific attributes that are part of the organizational name. For example, email *7*.

Performance Evaluation:

- Prompt P2_1:

  > 4. Organization: Extract the name of the organization offering the position.

  - **Jaccard Similarity: 0.699**
  - **Strengths:** Provides a clear and straightforward structure for extraction. Emphasizes capturing information exactly as it appears, which helps maintain consistency.
  - **Weaknesses:** Lacks specific guidance on handling variations in organization names or subunits, which might lead to incomplete or less accurate extraction in cases where subunits or acronyms are used.

- Prompt P2_2:

  > 4. Organization: Extract the full name of the organization from the email content, including subunits and departments. Ensure the name is complete, maintains consistent naming conventions, and avoids abbreviations or acronyms. The organization name should match exactly how it is provided in the email, without any paraphrasing or omissions.

  - **Jaccard Similarity: 0.685**
  - **Strengths:** Provides detailed instructions on extracting the organization name, including considerations for subunits and avoiding abbreviations. Emphasizes capturing names exactly as provided, which is crucial for accurate extraction.
  - **Weaknesses:** The length and complexity might make it harder to follow, potentially leading to errors or missed details. While comprehensive, the prompt may still be prone to inconsistencies if there are variations in naming conventions or abbreviations.

- Prompt P2_3:

  > 4. Organization: Extract the organization's full name from the email content, including subunits and departments. Ensure the name is complete, maintains consistent naming conventions, and avoids abbreviations or acronyms. The organization name should match exactly how it is provided in the email, without any paraphrasing or omissions.

  - **Jaccard Similarity: 0.824**
  - **Strengths:** Highly detailed and specific, providing examples to clarify the expected format and content. The additional examples help guide accurate extraction and ensure consistency.
  - **Weaknesses:** The length and complexity of the prompt may cause potential confusion or errors if not followed precisely. The comprehensive nature might be overwhelming, leading to inconsistencies in adherence to the provided instructions.

4.2.5 Extraction of the Group/Department Name

The table below shows a few examples where the group or department names are not correctly predicted with prompt P2_1. The complete table is in Appendix A, A.15.

No	Group	Group (pr)
1	People and Information Research Team (PIRE-T)	Dept. of Computer Science
2	TETIS Unit (Territories, Environment, Remote Detection and Spatial Information)	TETIS Unit
3	SISRG (Strathclyde iSchool Research Group)	Department of Computer and Information Sciences
4	Science, Engineering and Technology Group	Department of Computer Science

Table 4.8: Emails with incorrectly predicted group -P2_1

Error Analysis:

1. **Ambiguity among Research Group and Department**: There is confusion between the group and department, therefore; the model sometimes fails to extract the group or returns incorrect predictions. For example, emails *1, 3, 4*.
2. **Abbreviations of Groups**: The model returned the abbreviated form of groups instead of full name, such as email *2*.

Performance Evaluation:

- Prompt P2_1:

  > 5. Group or Department: Specify the group or department related to the position, if mentioned.

  - **Jaccard Similarity:** 0.451
  - **Strengths**: The prompt is straightforward, making it easy for the model to understand the task.
  - **Weaknesses**: The simplicity of the prompt leads to incorrect predictions when multiple groups are mentioned.

- Prompt P2_2:

  > 5. Group or Department: Extract the full and precise name of the specific group or department mentioned in the email. Utilize contextual clues such as units, research groups, or departments to ensure the extracted name is accurate and complete. The name should match exactly how it is provided in the email, without any paraphrasing or omissions.

  - **Jaccard Similarity:** 0.583

  - **Strengths**: The prompt explicitly addresses the inclusion of multiple groups, guiding the model to provide a more accurate extraction.
  - **Weaknesses**: Despite detailed instructions, the prompt might still struggle with ambiguities related to broad terms and the level of detail. The requirement to match detail levels exactly could make it more difficult for the model to perform well.

- Prompt P2_3:

  > 5. Group or Department: Extract the full and precise name of the specific group or department mentioned in the email. Utilize contextual clues such as units, research groups, or departments for identifying groups. The name should match exactly how it is provided in the email, without any paraphrasing or omissions.

  - **Few-shot prompting:** Provided examples to guide the model.
  - **Jaccard Similarity:** 0.576
  - **Strengths**: The few-shot examples help clarify the expected output, providing the model with context and improving its understanding and performance.
  - **Weaknesses**: Despite the guidance from examples, the added complexity might have confused the model, leading to a slight decrease in performance. The model might overfit the provided examples, making it less flexible in generalizing to new, unseen prompts.

The following table summarizes the evaluation results for the evaluated job position emails.

Category	Prompt	Data	Precision	Recall	F1 score	Jaccard
contact-person	P2_1	dev	0.733	0.815	0.772	0.802
position	P2_1	dev	0.241	0.233	0.237	0.391
work-area	P2_1	dev	0.200	0.207	0.203	0.342
organization	P2_1	dev	0.500	0.500	0.500	0.699
group	P2_1	dev	0.435	0.333	0.377	0.451
contact-person	P2_2	dev	0.621	0.667	0.643	0.663
position	P2_2	dev	0.100	0.100	0.100	0.289
work-area	P2_2	dev	0.148	0.138	0.143	0.379
organization	P2_2	dev	0.433	0.433	0.433	0.685
group	**P2_2**	dev	**0.552**	**0.571**	**0.561**	**0.583**
contact-person	**P2_3**	dev	**0.767**	**0.793**	**0.78**	**0.838**
position	**P2_3**	dev	**0.433**	**0.433**	**0.433**	**0.559**
work-area	**P2_3**	dev	**0.207**	**0.207**	**0.207**	**0.404**
organization	**P2_3**	dev	**0.700**	**0.700**	**0.700**	**0.824**
group	P2_3	dev	0.483	0.500	0.491	0.576
contact-person	P2_1	test	0.737	0.778	0.757	0.733
position	P2_1	test	0.3	0.3	0.3	0.428
work-area	P2_1	test	0.1	0.1	0.1	0.373
organization	P2_1	test	0.75	0.75	0.75	0.814
group	P2_1	test	0.61	0.625	0.617	0.542
contact-person	P2_2	test	0.632	0.706	0.667	0.646
position	P2_2	test	0.3	0.3	0.3	0.443
work-area	P2_2	test	0.105	0.1	0.103	0.304
organization	P2_2	test	0.65	0.65	0.65	0.775
group	P2_2	test	0.526	0.667	0.588	0.525
contact-person	P2_3	test	0.73	0.745	0.737	0.81
position	P2_3	test	0.41	0.41	0.41	0.49
work-area	P2_3	test	0.19	0.2	0.194	0.39
organization	P2_3	test	0.732	0.767	0.749	0.812
group	P2_3	test	0.59	0.601	0.595	0.552

Table 4.9: Model Performance Evaluation of Job Announcement Emails

Chapter 5

Conclusion and Future Work

This thesis aims to enhance the Information Retrieval Anthology by incorporating valuable data extracted from mailing lists, such as the SIGIR archive. Through the use of advanced natural language processing techniques, particularly leveraging state-of-the-art language models, the research focused on the extraction of significant insights from unstructured email data.

The results underscore the potential of the methodology to enrich the IR Anthology, providing researchers with more comprehensive and relevant resources. The integration of additional information from mailing lists into the IR Anthology is a primary objective, and this study demonstrates a feasible pathway toward achieving this goal. By enhancing the database with diverse and up-to-date content, the Anthology becomes a more valuable tool for the research community.

However, challenges remain, particularly in improving the accuracy of information extraction models, especially in handling ambiguous terms and abbreviations. These issues are highlighted in the evaluation phase and pointed towards areas where further refinement is needed.

Looking forward, the main focus of future work should be on seamlessly integrating the extracted information with the IR Anthology. This involves not only improving the accuracy of data extraction but also developing robust systems for data integration and retrieval within the Anthology framework. Additionally, exploring ways to incorporate other forms of unstructured data and expanding the dataset beyond the current mailing lists will be crucial for broadening the scope and utility of the IR Anthology. Future work should also focus on fine-tuning the current models to improve their precision and recall. By optimizing these models, we can achieve more accurate and reliable data extraction. Additionally, further exploration of fine-tuning techniques, such as transfer learning and hyperparameter optimization, will be critical in enhancing model performance.

In conclusion, this thesis contributes to the ongoing development of the IR Anthology by providing the methodology for extracting and integrating new data sources. The successful implementation of this approach promises to significantly enhance the utility of the Anthology, making it a richer and more accessible channel for researchers in the field of information retrieval. The work lays a strong foundation for future developments, aiming towards a more comprehensive and integrated knowledge base.

Appendix A

Incorrect Predictions

No	Topics of Interest	Topics of Interest(pr)
1	['information (seeking and searching) behaviour (ib)', 'human computer interaction (hci)', 'information retrieval (ir)']	['user-centered approaches to the design and evaluation of systems for information access, retrieval, and use', 'information (seeking and searching) behaviour (ib)', 'human computer interaction (hci)', 'information retrieval (ir)']
2	['total recall problem', 'diagnostic test accuracy (dta) reviews']	['technologically assisted reviews in empirical medicine', 'systematic review articles', 'total recall problem', 'automation in systematic review process', 'diagnostic test accuracy reviews', 'abstract and title screening']
3	['innovative and risky information access and retrieval system ideas', 'systems-building experience and insight', 'resourceful experimental studies', 'provocative position statements', 'new application domains']	['papers', 'prototypes', 'abstracts', 'open problems in ir']
4	['future directions of information access', 'early research such as pilot studies, presenting challenges and future opportunities, conceptual and theoretical work, and contributions from doctoral work']	['future directions of information access', 'young researchers', 'early research', 'pilot studies', 'challenges and future opportunities', 'conceptual and theoretical work', 'contributions from doctoral work', 'information retrieval', 'interaction and usage']

Continued on next page

Table A.1 – continued from previous page

No	Topics of Interest	Topics of Interest(pr)
5	['new or improved models of relevance', 'ranking', 'representation', 'information needs', 'evaluation', 'formal frameworks', 'new search paradigms', 'methods from related disciplines']	['new or improved models of relevance', 'ranking', 'representation', 'information needs and evaluation', 'formal frameworks in information retrieval', 'defining new tasks and search paradigms', 'applying methods from related disciplines']
6	[]	['tutorials', 'keynotes', 'panel session on evaluation initiatives', 'mirrors special session', 'regular paper sessions', 'concerts', 'social program']
7	[]	['trec tracks for trec 2016']
8	['web search and data mining']	['web search', 'data mining']

Table A.1: Emails with incorrectly predicted topics of interest-P1_1

No	Topics of Interest	Topics of Interest(pr)
1	['new theory and models for urban information retrieval and big data computing', 'ultra-high efficiency compression, coding and transmission for urban data', 'system architectures, scalability and efficiency', 'personalization, recommendation and filtering of urban data', 'deep learning and cloud computing for urban informatics', 'information extraction, knowledge representation, and reasoning over urban data', 'information retrieval access techniques for urban informatics', 'use of social media in urban informatics', 'spatial, temporal, and graph data mining for urban informatics', 'filtering, time-sensitive and real-time search', 'novel applications of information retrieval and big data approaches in urban informatics (e.g., smart cities, healthcare, energy, transportation, safety, crime, search and retail)', 'interaction, access, and visualization of urban data', 'user-based evaluation, crowdsourcing-based evaluation, click models, interaction models in urban informatics', 'ethics, privacy and security in urban informatics']	['new theory and models for urban information retrieval and big data computing', 'ultra-high efficiency compression, coding and transmission for urban data', 'system architectures, scalability and efficiency', 'personalization, recommendation and filtering of urban data', 'deep learning and cloud computing for urban informatics', 'information extraction, knowledge representation, and reasoning over urban data', 'information retrieval access techniques for urban informatics', 'use of social media in urban informatics', 'spatial, temporal, and graph data mining for urban informatics', 'filtering, time-sensitive and real-time search', 'novel applications of information retrieval and big data approaches in urban informatics (e.g., smart cities, healthcare, energy, transportation, safety, crime, search and retail)', 'interaction, access, and visualization of urban data', 'user-based evaluation, crowdsourcing-based evaluation, click models, interaction models in urban informatics', 'ethics, privacy, and security in urban informatics']
2	['total recall problem', 'diagnostic test accuracy (dta) reviews']	['diagnostic test accuracy (dta) reviews']

Continued on next page

Table A.2 – continued from previous page

No	Topics of Interest	Topics of Interest(pr)
3	['user aspects including information interaction, contextualisation, personalisation, simulation, characterisation, and information behaviours', 'system aspects including retrieval and recommendation algorithms, machine learning, deep learning, content representation, natural language processing, system architectures, and efficiency methods', 'applications such as search and recommender systems, web and social media apps, domain specific search (professional, bio, chem, etc.), novel interfaces, intelligent search agents/bots, and related innovative search tools', 'evaluation research including new measures and novel methods for the measurement and evaluation of users, systems and/or applications']	['user aspects', 'system aspects', 'applications', 'evaluation research', 'ehealth', 'deeplearning', 'education ir']
4	['future directions of information access', 'early research such as pilot studies, presenting challenges and future opportunities, conceptual and theoretical work, and contributions from doctoral work']	['the early research such as pilot studies', 'presenting challenges and future opportunities', 'conceptual and theoretical work', 'contributions from doctoral work']
5	['new or improved models of relevance', 'ranking', 'representation', 'information needs', 'evaluation', 'formal frameworks', 'new search paradigms', 'methods from related disciplines']	['new or improved models of relevance', 'ranking', 'representation', 'information needs', 'evaluation', 'define new tasks', 'develop new search paradigms', 'apply methods from related disciplines']

Table A.2: Emails with incorrectly predicted topics of interest-P1_2

No	Contact Person	Contact Person(pr)
1	prof. iryna gurevych, dr. ivan habernal	dr. ivan habernal
2		unspecified
3		iadh.ounis@glasgow.ac.uk
4	professor ian ruthven (head of sisrg) and/or dr martin halvey (director postgraduate teaching)	dr martin halvey

Continued on next page

Table A.3 – continued from previous page

No	Contact Person	Contact Person(pr)
5	prof. michael twidale, dr. jodi schneider	dr. jodi schneider
6	associate professor isabelle augenstein	isabelle augenstein
7	prof. brian corcoran or prof. andy way	prof. brian corcoran

Table A.3: Emails with incorrectly predicted contact person -P2_1

No	Contact Person	Contact Person(pr)
1	prof. iryna gurevych, dr. ivan habernal	
2	j. stephen downie, phd	j. stephen downie, associate dean for research
3	dr. damien sileo or prof. dr. marie-francine moens	dr. damien sileo
4		department of computer science
5		iadh ounis
6	professor ian ruthven (head of sisrg) and/or dr martin halvey (director postgraduate teaching)	dr martin halvey
7	prof. michael twidale, dr. jodi schneider	prof. michael twidale, phd program director
8	prof. michael twidale	halil kilicoglu
9		dr. rachel charlotte smith
10	associate professor isabelle augenstein	isabelle augenstein
11	prof. brian corcoran or prof. andy way	prof. brian corcoran, acting executive dean, faculty of engineering and computing
12	raja jurdak	associate professor laurianne sitbon

Table A.4: Emails with incorrectly predicted contact person -P2_2

No	Contact Person	Contact Person(pr)
1	dr. damien sileo or prof. dr. marie-francine moens	dr. damien sileo, prof. dr. marie-francine moens
2	professor ian ruthven (head of sisrg) and/or dr martin halvey (director postgraduate teaching)	professor ian ruthven, dr. martin halvey
3	prof. michael twidale, dr. jodi schneider	dr. jodi schneider
4		reference person for each topic

Continued on next page

Table A.5 – continued from previous page

No	Contact Person	Contact Person(pr)
5	associate professor isabelle augenstein	isabelle augenstein, phd
6	prof. brian corcoran or prof. andy way	prof. brian corcoran, prof. andy way
7	raja jurdak	raja jurdak, dr. yan zhang,ph.d

Table A.5: Emails with incorrectly predicted contact person -P2_3

No	Position	Position(pr)
1	PhD position	PhD positions in Computer Science
2	Professorship	Professor of Machine Learning / Data Science
3	Research scientists	Research Scientists (m/f/d)
4	Tenure-track faculty	Tenure-track faculty in Digital Archives, Preservation, and/or Curation (open rank)
5	PhD student	PhD position: Methods for Review Processes Using AI and ML
6	PhD program	Information Sciences PhD Program
7	Research position	Research Position for Postdoctoral or Predoctoral Scientist on Explainable AI in Clinical Reports
8	Researcher	Researcher in Data Science and Modelling
9	PhD position	PhD Position in Responsible Natural Language Processing and Data Management for Mental Health
10	Lecturer or assistant professor	Academic Position at the Rank of Lecturer or Assistant Professor
11	Fellowships	Visiting Scientists/Professorships @ Leibniz University of Hannover / L3S & CISPA
12	Lecturer/ Senior Lecturer/ Reader position	Lecturer/ Senior Lecturer/ Reader
13	Senior lecturer/lecturer	Senior Lecturer/Lecturer in Information Behaviour
14	PhD students	Fully Funded PhD Students in Information Sciences at the School of Information Sciences (iSchool), University of Illinois at Urbana-Champaign
15	Doctoral program	
16	Junior professor	Junior Professor in Natural Language Processing and Multimedia Interaction

Continued on next page

Table A.6 – continued from previous page

No	Position	Position(pr)
17	Junior professor	Academic Vacancy in the Area of Natural Language Processing and Multimedia Interaction
18	PhD positions	PhDs on the Topics Indicated Below
19	PhD positions	Prototyping New Professional Roles and Design Practices for the Digital Society
20	PhD scholarships	PhD Scholarships in Natural Language Processing
21	Ph.D.	Ph.D. in Information Science Program
22	Full professor	Full Professor of Computing (Multimodal Information Systems)
23	Doctoral program	Doctor of Philosophy Program in Information Science

Table A.6: Emails with incorrectly predicted position -P2_1

No	Position	Position (pr)
1	PhD position	PhD students
2	Professorship	Professor of Machine Learning / Data Science
3	Research scientists	Research Scientists (m/f/d)
4	Tenure-track faculty	Tenure-track faculty in Digital Archives, Preservation, and/or Curation (open rank)
5	Assistant or Associate Professor	Associate Professor
6	PhD student	PhD position in Methods for Review Processes Using AI and ML
7	PhD program	Information Sciences PhD Program
8	Research position	Postdoctoral or Predoctoral Scientist
9	Researcher	Researcher in Data Science and Modelling
10	PhD position	Fully-funded PhD on Responsible Natural Language Processing and Data Management for Mental Health
11	Lecturer or Assistant Professor	Lecturer or Assistant Professor in the Department of Computer Science
12	Fellowships	Visiting Scientists/Professorships
13	Lecturer/ Senior Lecturer/ Reader position	Lecturer/ Senior Lecturer/ Reader
14	Senior Lecturer/Lecturer	Director Postgraduate Teaching
15	Assistant Professor	Assistant Professor in Information Systems

Continued on next page

Table A.7 – continued from previous page

No	Position	Position (pr)
16	PhD students	PhD Program Director
17	Doctoral program	PhD Program Director
18	Junior Professor	Junior Professor in Natural Language Processing and Multimedia Interaction
19	Junior Professor	Full-time Tenure-track Academic Vacancy
20	PhD positions	PhDs on the Topics Indicated Below
21	PhD positions	Design Anthropology for Sustainable Human-Machine Relations
22	PhD scholarships	Associate Professor

Table A.7: Emails with incorrectly predicted position -P2_2

No	Position	Position(pr)
1	PhD position	call for PhD positions
2	Professorship	Professor of Machine Learning / Data Science
3	Research scientists	Research scientist
4	PhD program	Ph.D. program
5	Lecturer or assistant professor	Tenure-track academic position at the rank of Lecturer or Assistant Professor
6	Fellowships	Leibniz AI & Security Fellowship Program 2022
7	Lecturer/ Senior Lecturer/ Reader position	Lecturer/ Senior Lecturer/ Reader
8	PhD students	Dr.
9	Doctoral program	PhD Program Director
10	PhD positions	PhDs
11	PhD positions	15 fully funded PhD positions
12	Ph.D.	Ph.D. in Information Science Program
13	Full professor	Full Professor of Computing (Multimodal Information Systems)
14	Doctoral program	Ph.D. program

Table A.8: Emails with incorrectly predicted position -P2_3

No	Work Area	Work Area (pr)
1	AI for data access and integration, semantic technologies, explainable and strategic AI, neuro-symbolic AI for business-process analysis, AI for cybersecurity, process mining on object networks, graph-data management, machine learning for knowledge graphs, temporal reasoning and verification, graph data management with linear algebra, process mining on object networks, ontology-driven belief propagation for cybersecurity, learning mappings in virtual knowledge graphs, transforming and explaining data and knowledge, strategy and explainability for knowledge bases, neuro-symbolic artificial intelligence for business process analysis	Knowledge representation in artificial intelligence
2	Machine learning, data science, explainable AI (XAI), responsible AI or trustworthy AI	Machine learning / data science
3	Privacy-preserving natural language processing, security and privacy in artificial intelligence (SENPAI)	Privacy-preserving natural language processing
4	Digital archives, digital preservation, digital curation, born digital preservation and data curation, cultural heritage and collections as data, critical archival and data studies, software and platform studies, community-centered digital practices and preservation, critical approaches to metadata, digital accessibility, digital sustainability, digital preservation infrastructures, tools, and policies	Information sciences
5	Data science, AI, machine learning, and/or information retrieval	Data science, AI, machine learning, information retrieval
6	Systematic reviews (of intervention, diagnostic accuracy, and prognosis studies) and methodological research on evidence synthesis in the medical domain	Medical domain
7	Natural language processing, machine learning, explainable AI	Natural language processing and clinical reports

Continued on next page

Table A.9 – continued from previous page

No	Work Area	Work Area (pr)
8	Responsible natural language processing and data management for mental health	Natural language processing and data management for mental health
9	Artificial intelligence, data science, information security	Artificial intelligence and data science or information security
10	Information systems, bioinformatics and human computer interaction	Information systems
11	Scientific information, scholarly communication, controversies within science, potential sources of bias in scientific research, confidence in applying science to public policy	Information sciences
12	Archives & digital curation, artificial intelligence & data science, digital humanities, health, medical & bioinformatics, human-computer interaction, information organization & access, information, culture & society, library technologies & services, science of science, social computing & computational social science	Information sciences
13	Natural language processing and multimedia interaction	Science, engineering, and technology
14	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet: interconnection of quantum and legacy-internet networks, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications
15		Information
16	Design, anthropology, media studies, science and technology studies, design informatics, human-data/AI interaction, artificial intelligence, machine learning	Design
17	Theoretical and applied machine learning, natural language processing and understanding, information retrieval, and medical image analysis	Natural language processing

Continued on next page

Table A.9 – continued from previous page

No	Work Area	Work Area (pr)
18	Multimodal information systems, language technology, machine translation, personalisation, VR/AR, privacy and ethics	Multimodal information systems
19	Spatiotemporal machine learning/artificial intelligence	Spatiotemporal machine learning/artificial intelligence algorithms
20	Archival studies, biodiversity information and culture, computer-mediated communication, critical information and library science, cultural analytics, cultural heritage informatics, data and text mining, digital and computational humanities, digital curation, digital libraries, discourse and multimodality, documentation, eScience, gender and technology, history of the book, readership, and publishing, knowledge organization (KO) (classification, ontologies, metadata), knowledge sharing, natural language processing, information retrieval, online communities, philosophy of information and information ethics, records management, social informatics, social media mining, technology and diversity, text encoding and digital editions	Information science

Table A.9: Emails with incorrectly predicted work area -P2_1

No	Work Area	Work Area (pr)
1	AI for data access and integration, semantic technologies, explainable and strategic AI, neuro-symbolic AI for business-process analysis, AI for cybersecurity, process mining on object networks, graph-data management, machine learning for knowledge graphs, temporal reasoning and verification, graph data management with linear algebra, process mining on object networks, ontology-driven belief propagation for cybersecurity, learning mappings in virtual knowledge graphs, transforming and explaining data and knowledge, strategy and explainability for knowledge bases, neuro-symbolic artificial intelligence for business process analysis	AI for data access and integration, semantic technologies, explainable and strategic AI, neuro-symbolic AI for business-process analysis, AI for cybersecurity, process mining on object networks, graph-data management, machine learning for knowledge graphs, temporal reasoning and verification
2	Machine learning, data science, explainable AI (XAI), responsible AI or trustworthy AI	Explainable AI (XAI), responsible AI or trustworthy AI
3	Privacy-preserving natural language processing, security and privacy in artificial intelligence (SENPAI)	Security and privacy in artificial intelligence
4	Digital archives, digital preservation, digital curation, born digital preservation and data curation, cultural heritage and collections as data, critical archival and data studies, software and platform studies, community-centered digital practices and preservation, critical approaches to metadata, digital accessibility, digital sustainability, digital preservation infrastructures, tools, and policies	
5	Data science, AI, machine learning, and/or information retrieval	Data science, AI, machine learning, information retrieval
6	Systematic reviews (of intervention, diagnostic accuracy, and prognosis studies) and methodological research on evidence synthesis in the medical domain	Medical domain
7	Information sciences	Information sciences PhD program
8	Natural language processing, machine learning, explainable AI	Natural language processing

Continued on next page

Table A.10 – continued from previous page

No	Work Area	Work Area (pr)
9	Data science and modelling	Animal health monitoring, food safety monitoring
10	Responsible natural language processing and data management for mental health	Natural language processing, data management, neural pipelines, cross-lingual/domain adaptation, data integration
11	Artificial intelligence, data science, information security	Artificial intelligence in Hannover
12	Information systems, bioinformatics and human computer interaction	Bioinformatics, human computer interaction
13	Scientific information, scholarly communication, controversies within science, potential sources of bias in scientific research, confidence in applying science to public policy	Scientific information, scholarly communication, controversies within science, potential sources of bias in scientific research, confidence in applying science to public policy, knowledge representation, text and data analytics, news analytics, argumentation analysis, document analysis, qualitative analysis, user-centered design, mixed methods, public policy, public health, libraries, journalism, publishing, citizen science, information services, life sciences research
14	Archives & digital curation, artificial intelligence & data science, digital humanities, health, medical & bioinformatics, human-computer interaction, information organization & access, information, culture & society, library technologies & services, science of science, social computing & computational social science	Information science
15	Natural language processing and multimedia interaction	Natural language processing, multimedia interaction
16	Natural language processing and multimedia interaction	Natural language processing, multimedia interaction

Continued on next page

Table A.10 – continued from previous page

No	Work Area	Work Area (pr)
17	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet: interconnection of quantum and legacy-internet networks, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications
18		Information science
19	Design, anthropology, media studies, science and technology studies, design informatics, human-data/AI interaction, artificial intelligence, machine learning	Design anthropology
20	Theoretical and applied machine learning, natural language processing and understanding, information retrieval, and medical image analysis	Natural language processing, understanding, information retrieval, medical image analysis

Table A.10: Emails with incorrectly predicted work area -P2_2

No	Work Area	Work Area (pr)
1	AI for data access and integration, semantic technologies, explainable and strategic AI, neuro-symbolic AI for business-process analysis, AI for cybersecurity, process mining on object networks, graph-data management, machine learning for knowledge graphs, temporal reasoning and verification, graph data management with linear algebra, process mining on object networks, ontology-driven belief propagation for cybersecurity, learning mappings in virtual knowledge graphs, transforming and explaining data and knowledge, strategy and explainability for knowledge bases, neuro-symbolic artificial intelligence for business process analysis	Computer science

Continued on next page

Table A.11 – continued from previous page

No	Work Area	Work Area (pr)
2	Machine learning, data science, explainable AI (XAI), responsible AI or trustworthy AI	Machine learning / data science, explainable AI, responsible AI, trustworthy AI
3	Privacy-preserving natural language processing, security and privacy in artificial intelligence (SENPAI)	Security and privacy in artificial intelligence, privacy-preserving natural language processing, digital mental health, privacy-aware domain-adaptive medical NLP, protecting privacy and sensitive information in texts, large language models for clinical applications, machine learning for natural language processing, differential privacy
4	Digital archives, digital preservation, digital curation, born digital preservation and data curation, cultural heritage and collections as data, critical archival and data studies, software and platform studies, community-centered digital practices and preservation, critical approaches to metadata, digital accessibility, digital sustainability, digital preservation infrastructures, tools, and policies	Digital archives, digital preservation, digital curation
5	Data science, AI, machine learning, and/or information retrieval	Data science, AI, machine learning, information retrieval
6	Systematic reviews (of intervention, diagnostic accuracy, and prognosis studies) and methodological research on evidence synthesis in the medical domain	Evidence synthesis, information retrieval, computer science
7	Natural language processing, machine learning, explainable AI	Explainable AI, natural language processing
8	Responsible natural language processing and data management for mental health	Responsible natural language processing, data management, mental health
9	Artificial intelligence, data science, information security	Artificial intelligence and data science, information security
10	Information behaviour	Information behaviour, information engagement, digital cultural heritage, interactive information retrieval
11	Information systems, bioinformatics and human computer interaction	Information systems, bioinformatics, human computer interaction

Continued on next page

Table A.11 – continued from previous page

No	Work Area	Work Area (pr)
12	Scientific information, scholarly communication, controversies within science, potential sources of bias in scientific research, confidence in applying science to public policy	Information sciences, information quality
13	Archives & digital curation, artificial intelligence & data science, digital humanities, health, medical & bioinformatics, human-computer interaction, information organization & access, information, culture & society, library technologies & services, science of science, social computing & computational social science	PhD information science program
14	Natural language processing and multimedia interaction	Natural language processing and multimedia interaction
15	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications	Human-centric artificial intelligence, decentralised AI for resource-constrained edge systems, quantum internet: interconnection of quantum and legacy-internet networks, resource allocation and slicing in MEC systems for latency-sensitive IoT applications, AI-based decision support systems for smart healthcare applications
16		Information
17	Theoretical and applied machine learning, natural language processing and understanding, information retrieval, and medical image analysis	Natural language processing, information retrieval, medical image analysis
18	Information science	Information science, information systems, information behavior, digital youth, social and cultural studies, digital humanities, information policy
19	Multimodal information systems, language technology, machine translation, personalization, VR/AR, privacy and ethics	Multimodal information systems
20	Spatiotemporal machine learning/artificial intelligence	Spatiotemporal machine learning/artificial intelligence algorithms

Continued on next page

Table A.11 – continued from previous page

No	Work Area	Work Area (pr)
21	Archival studies, biodiversity information and culture, computer-mediated communication, critical information and library science, cultural analytics, cultural heritage informatics, data and text mining, digital and computational humanities, digital curation, digital libraries, discourse and multimodality, documentation, eScience, gender and technology, history of the book, readership, and publishing, knowledge organization (KO) (classification, ontologies, metadata), knowledge sharing, natural language processing, information retrieval, online communities, philosophy of information and information ethics, records management, social informatics, social media mining, technology and diversity, text encoding and digital editions	

Table A.11: Emails with incorrectly predicted work area -P2_3

No	Organization	Organization (pr)
1	University of Illinois at Urbana-Champaign	University of Illinois, Urbana-Champaign
2	Department of Computer Science, Boise State University	Boise State University
3	University Medical Center Utrecht	Cochrane Netherlands
4	CIRAD (French Agricultural Research Centre for International Development)	CIRAD
5	Leibniz University of Hannover / L3S & CISPA	Leibniz AI & Security Fellowships
6	University of Glasgow	The University of Glasgow
7	Computer Science Departfaculty of Sciences (ULisboa)	Computer Science Department of the Faculty of Sciences (ULisboa)
8	University of Illinois at Urbana-Champaign	University of Illinois
9	KU Leuven	KU Leuven (Belgium)
10	IIT-CNR, Pisa, Italy	The Ubiquitous Internet Research Unit @ IIT-CNR, Pisa, Italy
11	San José State University	San José State University

Continued on next page

Table A.12 – continued from previous page

No	Organization	Organization (pr)
12	Marie Skłodowska-Curie European Training Network DCODE	Amsterdam University of Applied Sciences
13	University of Texas at Austin	The University of Texas at Austin's School of Information (iSchool)
14	University of Oklahoma	School of Library and Information Studies (SLIS) at the University of Oklahoma
15	Indiana University Bloomington	Indiana University

Table A.12: Emails with incorrectly predicted organization -P2_1

No	Organization	Organization (pr)
1	Free University of Bozen-Bolzano	Free University of Bozen-Bolzano, Italy
2	University of Illinois at Urbana-Champaign	University of Illinois, Urbana-Champaign
3	Department of Computer Science, Boise State University	Boise State University
4	iSchool @ Illinois	School of Information Sciences (iSchool) at the University of Illinois Urbana-Champaign
5	CIRAD (French Agricultural Research Centre for International Development)	CIRAD
6	Leibniz University of Hannover / L3S & CISPA	Leibniz AI & Security Fellowships
7	University of Glasgow	The University of Glasgow
8	Computer Science Departfaculty of Sciences (ULisboa)	Computer Science Department of the Faculty of Sciences (ULisboa)
9	University of Illinois at Urbana-Champaign	School of Information Sciences (iSchool), University of Illinois at Urbana-Champaign
10	University of Illinois at Urbana-Champaign	School of Information Sciences, University of Illinois at Urbana-Champaign
11	KU Leuven	KU Leuven (Belgium), Faculty of Engineering Science, Department of Computer Science
12	San José State University	San José State University
13	Marie Skłodowska-Curie European Training Network DCODE	University of Aarhus
14	Max Planck Institutes	Max Planck Institutes for Informatics, Software Systems, and Security and Privacy

Continued on next page

Table A.13 – continued from previous page

No	Organization	Organization (pr)
15	University of Texas at Austin	The University of Texas at Austin
16	University of Oklahoma	The University of Oklahoma
17	Indiana University Bloomington	Indiana University

Table A.13: Emails with incorrectly predicted organization -P2_2

No	Organization	Organization (pr)
1	Department of Computer Science, Boise State University	Boise State University
2	iSchool @ Illinois	University of Illinois Urbana-Champaign
3	CIRAD (French Agricultural Research Centre for International Development)	CIRAD
4	Leibniz University of Hannover / L3S & CISPA	Leibniz University of Hannover
5	Computer Science Departfaculty of Sciences (ULisboa)	Computer Science Department of the Faculty of Sciences (ULisboa)
6	IIT-CNR, Pisa, Italy	IIT-CNR
7	San José State University	San José State University
8	Marie Skłodowska-Curie European Training Network DCODE	DCODE - Marie Skłodowska-Curie European Training Network
9	Max Planck Institutes	Max Planck Society

Table A.14: Emails with incorrectly predicted organization -P2_3

No	Group	Group (pr)
1	Ubiquitous Knowledge Processing (UKP) Lab, Trustworthy Human Language Technologies (TrustHLT)	Trustworthy Human Language Technologies (TrustHLT)
2	People and Information Research Team (PIRE-T)	Dept. of Computer Science
3	Cochrane Netherlands	
4	School of Information Sciences (iSchool)	School of Information Sciences
5	TETIS Unit (Territories, Environment, Remote Detection and Spatial Information)	TETIS Unit
6	Faculty of Electrical Engineering and Computer Science, L3S Research Center, CISPA / Center for Information Security	Faculty of Electrical Engineering and Computer Science
7	SISRG (Strathclyde iSchool Research Group)	Department of Computer and Information Sciences

Continued on next page

Table A.15 – continued from previous page

No	Group	Group (pr)
8	Computer Science Department	
9	School of Information Sciences (iSchool)	School of Information Sciences
10	Science, Engineering and Technology Group	Department of Computer Science
11	Science, Engineering and Technology Group	Department of Computer Science
12	The Ubiquitous Internet Research Unit	
13	School of Information	
14	DCODE	Digital Life Centre
15	School of Information (iSchool)	
16	School of Library and Information Studies (SLIS)	
17	DCU School of Computing, ADAPT Centre for Digital Content Technology	School of Computing
18	Collaboration Between QUT and CSIRO's Future Science Platform on Artificial Intelligence and Machine Learning (MLAI FSP)	CSIRO's Future Science Platform on Artificial Intelligence and Machine Learning (MLAI FSP)
19	Department of Information and Library Science	

Table A.15: Emails with incorrectly predicted group -P2_1

No	Group	Group (pr)
1	Ubiquitous Knowledge Processing (UKP) Lab, Trustworthy Human Language Technologies (TrustHLT)	Department of Computer Science
2	Cochrane Netherlands	
3	School of Information Sciences (iSchool)	iSchool @ Illinois
4		Academic Medical Center (AMC-UVA), Data Science Center, Indelab
5	Faculty of Electrical Engineering and Computer Science, L3S Research Center, CISPA / Center for Information Security	Faculty of Electrical Engineering and Computer Science at Leibniz Universität Hannover
6	Information, Data and Analysis (IDA) Section of the School of Computing Science	School of Computing Science
7	SISRG (Strathclyde iSchool Research Group)	Department of Computer and Information Sciences

Continued on next page

Table A.16 – continued from previous page

No	Group	Group (pr)
8	School of Information Sciences (iSchool)	School of Information Sciences
9	Science, Engineering and Technology Group	Human-Computer Interaction Research Unit
10	Science, Engineering and Technology Group	Human-Computer Interaction Research Unit
11	DCODE	Department for Digital Design and Information Studies
12		Max Planck Institute for Security and Privacy
13	School of Information (iSchool)	iSchool
14	Collaboration Between QUT and CSIRO's Future Science Platform on Artificial Intelligence and Machine Learning (MLAI FSP)	School of Computer Science

Table A.16: Emails with incorrectly predicted group -P2_2

No	Group	Group (pr)
1	Ubiquitous Knowledge Processing (UKP) Lab, Trustworthy Human Language Technologies (TrustHLT)	Trustworthy Human Language Technologies (TrustHLT) at the Department of Computer Science
2	People and Information Research Team (PIRET)	Dept. of Computer Science
3	TETIS Unit (Territories, Environment, Remote Detection and Spatial Information)	TETIS Unit
4		Academic Medical Center (AMC-UVA), Data Science Center, Indelab
5	Faculty of Electrical Engineering and Computer Science, L3S Research Center, CISPA / Center for Information Security	Faculty of Electrical Engineering and Computer Science
6	Information, Data and Analysis (IDA) Section of the School of Computing Science	School of Computing Science
7	SISRG (Strathclyde iSchool Research Group)	Department of Computer and Information Sciences (CIS)
8	Computer Science Department	Computer Science Department of the Faculty of Sciences (ULisboa)
9	School of Information Sciences (iSchool)	School of Information Sciences
10	DCODE	

Continued on next page

Table A.17 – continued from previous page

No	Group	Group (pr)
11	Natural Language Understanding Research Group	Machine Learning Section at the Department of Computer Science, Faculty of Science
12		Max Planck Institute for Security and Privacy
13	DCU School of Computing, ADAPT Centre for Digital Content Technology	ADAPT Centre for Digital Content Technology
14	Collaboration Between QUT and CSIRO's Future Science Platform on Artificial Intelligence and Machine Learning (MLAI FSP)	School of Computer Science
15	Department of Information and Library Science	School of Informatics, Computing and Engineering

Table A.17: Emails with incorrectly predicted group -P2_3

Bibliography

Douglas E Appelt, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson. Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178, 1993.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL `https://arxiv.org/abs/2005.14165`.

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*, 2022.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR, 2023.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

OpenAI. Gpt-4 technical report, 2023. URL `https://www.openai.com/research/gpt-4`.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018. URL `https://arxiv.org/abs/1801.06146`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.