Distribution Agreement

Boxin Zhao April 25, 2024

Impact of Data Analysis on Nascent Natural Language Processing Tasks

by

Boxin Zhao

Jinho D. Choi
Adviser

Department of Computer Science

Jinho D. Choi

Adviser

Shun Yan Cheung

Committee Member

Gary Motley

Committee Member

2024

Impact of Data Analysis on Nascent Natural Language Processing Tasks

By

Boxin Zhao

Jinho D. Choi

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Computer Science

2024

Abstract

Impact of Data Analysis on Nascent Natural Language Processing Tasks

By Boxin Zhao

Understanding the importance of data analysis is essential for Natural Language Processing (NLP) research. While it is widely recognized that the better the data quality, the better the model performance, little to no effort has been made on quantifying the impact of data analysis in NLP research. For nascent NLP tasks, this judgement is even harder. This thesis presents a study of the influence of noisy dataset, falsely-targeted dataset, and existing unsuitale dataset on model performance and the impact that data analysis could make on these three types of incompetent datasets, respectively. Through fixing the noise labels in a noisy dataset, we have improved the model performance from 69\% to 75\% with the model structure unchanged; through re-pointing the falsely-targeted dataset to the application scenario, we worked out a deploy-able version of the model; and through creating a new dataset spanning over 1,000 application scenarios, the model trained on our dataset outperforms models trained on other datasets and zero-shot GPT. Our work has shown that data analysis could have a significant impact on nascent NLP tasks for all kinds of NLP data.

Impact of Data Analysis on Nascent Natural Language Processing Tasks

By

Boxin Zhao

Jinho D. Choi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Computer Science

2024

Acknowledgements

I would like to express my sincere gratitude to Dr. Jinho D. Choi for his invaluable guidance and mentorship throughout my journey in the realm of Natural Language Processing (NLP), which culminates in this honor thesis. His expertise and support have been instrumental in shaping my research experience.

My sincere appreciations also to Dr. Elaine Fisher, Prof. Steve Pittard, Dr. Rebecca Thomas, and Dr. Vicki Hertzberg for their genuine support and contributions to the resume classification project mentioned in this thesis.

I would like to extend my appreciation to Dr. James D. Finch, whose collaboration on the Dialogue State Generation project discussed in this thesis has been marked by kindness and patience. His contributions have greatly enriched the depth of this work.

I am grateful to Professor Gary Motley for his understanding, patience, and unwavering inspiration throughout my academic journey. His mentorship has been invaluable in shaping my perspective and approach to learning.

I am deeply indebted to Dr. Shun Yan Cheung for his inspiring lectures and steadfast academic advising, which have been instrumental in charting my academic and career path.

Finally, I would like to thank all my friends, Yiran Ma, Jiaxin Yang, Yi Yang, Peilin Wu, Chenyu Wang, Kevin Wu, Chen Xue, Jinghao Cai, Xinyi Guan, Qingyi Li, Mo Zhou, Yujia Xiang, and Tse Yuan Chan for the friendship throughout my years at college.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In NLP research, if the data is not of good quality, the results would not be good regardless of what fancy technique is applied. Therefore, for NLP researchers, it is essential to analyze the data before training the model.

Regardless of how the data was created, data analysis is essential for NLP research. Traditionally, data used to train NLP model is created by annotating each entry of the data by human. Therefore, data creation has always been labor-intensive and expensive. Furthermore, while human-annotated data are generally of good quality, human annotators could still make mistake. Therefore, it is important for researchers to analyze the created data using all kinds of data analysis techniques, such as sampled error analysis or inter-annotator agreement scores. More recently, with the help of Large Language Models (LLMs) like the Generative Pretrained Transformer (GPT), data creation has become much easier than before, as we could now prompt engineer a LLM to create good quality data for us. However, multiple research have shown that LLMs suffer from problems like hallucination and insufficient language understanding. Therefore, verifying the data through error analysis is also necessary for data created by LLMs.

While it is generally recognized that data quality is important, and that the better data we have, the better results we will achieve, to what extent do data quality matter in NLP research still remains unclear. Because the cost of creating and refining dataset is so high, understanding when to or when not to create new or refine data is very important for NLP researchers. For this reason, understanding the impact of data analysis is essential in balancing the cost and return in NLP research.

My work contributes to helping NLP researchers understand the impact of data analysis on nascent tasks, enabling NLP researchers to better balance the cost and the return of doing data analysis.

Good Data A good data is generally (a) clean, (b) helpful for model's performance on the task, and (c) closely simulate the real-world application scenario. Based on these three key factors of a good data, one could derive 3 types of incompetent data that are not ideal for training the model: datasets that are not clean, datasets that are not helpful for model's performance, datasets that does not simulate the real world scenario. Respectively, we name them noisy data, unsuitable data, and falsely-targeted data.

Noisy Data A dataset could contain much noise, and therefore could negatively influence the model performance. Noise could be false labels for classification models or wrong output sequence for sequence-to-sequence models. If a dataset is noisy, deep learning models would learn from these noisy present in the dataset and perform worse than expected.

Falsely-targeted Data A dataset the does not target the application scenario is a falsely-targeted dataset. Even if such a dataset is not noisy, the model trained on such a dataset could not work in the actual application to which it is deployed.

Existing Unsuitable Data For newly-proposed tasks, there are often plenty of existing datasets that works well for related tasks that have already been well-established. However, just because these existing datasets have been proven to work well for relevant tasks does not necessarily mean that they will work well for the nascent task being proposed. When the existing dataset does not suit the new task well, it is an existing unsuitable dataset.

1.1 Research Question

Based on the 3 types of incompetent datasets mentioned above, there are 3 research questions, respectively:

- How much does a noisy dataset influence model performance? To what extent could data analysis help to fix noisy data?
- In what way does a falsely-targeted dataset affect project outcome? What could be done to re-point a falsely-targeted dataset back to the target application?
- When the existing dataset is unsuitable for the nascent task proposed for the project, is it worth it to create new data for the nascent task?

To answer these questions, we have conducted two projects on 3 different tasks: competence-level classification, resume & job description matching, and dialogue state generation, focusing on noisy data, falsely-targeted data, and existing unsuitable data, respectively. Through analyzing the role of data analysis in mitigating the effect of these types of incompetent datasets, I formulate my thesis statement as the following section presents.

1.2 Thesis Statement

Through conducting data analysis when creating the data and when we analyze the model output in the 2 projects, we have significantly boosted model performance. Specifically,

- Fixing noise labels in a noisy data resulted in a significant improvement in model performance.
- Re-pointing the falsely-targeted dataset made the model work in the application to which it is deployed.
- Creating new data when existing dataset does not work resulted in a model better than those trained on existing datasets.

Chapter 2

Background

2.1 Transformers

In my work, I have used a couple of deep learning models in our works. Here, I will provide a brief overview of what these models are and what they are capable of doing.

2.1.1 Bidirectional Encoder Representations from Transformer

The Bidirectional Encoder Representations of Transformer (BERT) model [5] is a deep learning model based on the transformer model structure [25]. Building on top of the transformer model structure, the BERT model is trained through 2 steps: pretraining and fine-tuning. During pretraining, which is an unsupervised step, the BERT model gains a general understanding of the language structure; and during fine-tuning, which is supervised, the BERT models learns information for a specific downstream task, like learning the structure of resumes for resume classification tasks like competence-level classification. Once pretrained on a large corpus, the BERT model could be easily adapted to a downstream task through fine-tuning. This makes it very versatile. In our case, we have obtained the BERT model pretrained on a large wikipedia corpus from Huggingface[1], and then we fine-tuned the model using our labeled resume dataset or encoded job description dataset to have it do resume classification. Pretraining the BERT model requires two steps: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), both of which are detailed in the BERT paper [5].

2.1.2 Robustly Optimized BERT Pretraining Approach

Robustly Optimized BERT Pretraining Approach (RoBERTa) [15] is an advanced natural language processing (NLP) model. It represents a significant advancement over its predecessor, BERT (Bidirectional Encoder Representations from Transformers) [5]. RoBERTa essentially builds upon the foundational principles of BERT while introducing several key improvements aimed at enhancing its performance and robustness.

One of the primary enhancements in RoBERTa is its training methodology. While BERT was trained using a masked language model (MLM) objective, where random tokens in a sentence were masked and the model had to predict them, RoBERTa employs a more extensive training regimen. It removes the ”next sentence prediction” task from BERT's training and significantly increases the amount of training data and batch size. This approach allows RoBERTa to better leverage the available data and learn more effectively from it.

Additionally, RoBERTa incorporates various optimization techniques such as dynamic masking during training, larger batch sizes, and longer training times. These optimizations contribute to RoBERTa's superior performance across a wide range of NLP tasks compared to BERT. Overall, RoBERTa represents a refinement and optimization of the BERT architecture, resulting in a more robust and effective model for natural language understanding and generation tasks.

[1] https://huggingface.co/

2.1.3 Text-to-Text Transfer Transformer

Text-To-Text Transfer Transformer (T5) [21] is a large language model. It is built upon the Transformer architecture [25], which is a neural network model designed for processing sequential data, particularly natural language text.

T5 is unique in that it follows a "text-to-text" approach, where all NLP tasks are framed as text-to-text transformations. This means that both the inputs and outputs to the model are represented as text strings. For example, instead of providing a separate input question and document for question answering, T5 receives a single input text containing the concatenated question and document, and it generates the answer text as the output.

During pretraining, T5 is trained on a diverse range of text-to-text tasks, including machine translation, text summarization, language modeling, and more. This approach allows T5 to learn a unified representation of language that can be applied to a wide range of natural language understanding and generation tasks.

After pretraining, T5 can be fine-tuned on specific downstream tasks with labeled data, similar to other large language models. Fine-tuning allows T5 to adapt its learned representations to perform well on these specific tasks.

In summary, T5 is a versatile language model that adopts a unified text-to-text approach for natural language processing tasks. By training on diverse text-to-text transformations, T5 learns to understand and generate human-like text across a broad spectrum of NLP tasks.

2.1.4 Generative Pretrained Transformer

Generative Pre-trained Transformer (GPT) [20] is a type of large language model developed by OpenAI. It is based on the Transformer architecture [25], which is a deep learning model designed to process sequential data, particularly natural language text, by capturing long-range dependencies and contextual information effectively.

In essence, GPT is trained on a vast amount of text data using unsupervised learning techniques. During training, it learns to predict the next word in a sequence of text given the preceding words. This process is known as autoregressive language modeling. By pretraining on a diverse range of text data, GPT learns to understand the structure and nuances of human language.

Once pretrained, GPT can be fine-tuned on specific downstream tasks with labeled data, such as text classification, language translation, or text generation. Fine-tuning allows GPT to adapt its learned representations to perform well on these specific tasks.

Overall, GPT is a powerful and versatile language model capable of understanding and generating human-like text across a wide range of natural language processing tasks. For example, his brief description of GPT is a piece of output from a GPT model after I input “give me a brief, high-level explanation of what is gpt.”

2.2 Resume Classification & Screening

Regarding resume classification, a work in 2018 [19] suggests promising results of using Convolutional Neural Network (CNN) [11] with Word Embedding Base Approach to classify resumes by domains. A 2022 survey [1] provided a comprehensive study on the performance of doing resume classification using traditional Machine Learning and Natural Language Processing approaches, including K Nearest Neighbors [18], Naive Bayes [2], and many others. A 2023 research [16] pushes the study of classifying resumes by more detailed domains through the use of Ensemble Learning. However, by far the most relevant research in classifying resumes according to competence levels and matching resumes with job descriptions is still the previous work that have proposed the two tasks [12], making use of the context-aware transformer models.

Dataset-wise, there exists in public domain some small datasets for resume classification based on categories [6]. A resume classification study using CNN used a dataset of IT resumes collected from the internet [9]. To the best of our knowledge, our labeled resume dataset is the only existing dataset that works for competence-level classification. While there are also job datasets that features a couple of job descriptions [22], there are no resumes to be matches to these job descriptions. Our job description dataset is, again to our knowledge, the only job description dataset that comes with positive and negative resume examples, and therefore capable of training resume & job description models.

2.3 Dialogue State Generation

Dialogue State Tracking (DST) is a well-established task that has been tackled under the fully-supervised [27], few-shot [4], and zero-shot settings [8, 10, 26]. Because my thesis is on data analysis, instead of going into details of their modeling approach, I would like to focus a couple of datasets dedicated to DST and how their dataset compares with ours.

The most diverse existing dataset for DST, Schema Guided Dialogues (SGD), covers 16 domains and 214 slot types [23]. MultiWOZ covers 7 domains and 24 slot types [3, 28]. Compared to these existing datasets, our newly-generated dataset features 1,000 dialogue domains with silver dialogue state labels, and thus it is exceptionally capable for training especially DSG models.

Chapter 3

Competence-Level Classification and Resume & Job Description Matching

3.1 Task Definition

Two new tasks in the field of automated resume screening, competence-level classification and resume & job description matching, was proposed in 2020 [13]. The attempt of using context-aware transformer models to do competence-level classification has shown promising results.

Competence-Level Classification Competence-Level Classification is to classify the resume according to the candidate's competence-level, which is essentially a measure of how competent the candidate is.

In the experiment settings of my work, I examined the hiring process of Clinical Research Coordinator (CRC) at Emory Healthcare. There are 4 levels of CRC, being CRC1, CRC2, CRC3, and CRC4, where CRC1 is the least competent position with lowest requirements and CRC4 is the most competent with the highest requirements.

The competence-level classification model would take a resume as input, and output the CRC level to which the candidate belongs. If the candidate does not qualify for any of the CRC levels, the candidate will be classified as `NQ`.

Resume & Job Description Matching Resume & Job Description Matching is a binary classification task that classifies whether the resume matches the job description.

The Resume & Job Description Matching model would take in a resume and a job description as input, and output the choice of whether the resume matches the job description. If the model outputs `YES`, that suggest that input resume and the job description is a good match; if the model outputs `NO`, then that means the resume and the job description is not a good match.

3.2 Task Motivation

Companies and institutions often get an overwhelming number of candidates for each and every job openings they offer. Going through all the resumes could be very labor-intensive. Therefore, in the traditional hiring process, companies and institutions would only go through a small portion of resumes they have received, and would not look at the rest once they have hired enough employees they need for the job opening, which is typically 1 or 2. This unfairness and bias in the traditional hiring process could cause harm for both the employer and the employee. For the employer, since not all the resumes received are read, they might not have found the most capable and suitable people they need for the job. And for the employees, this biased hiring process is unfair to people who applied for the job but do not get their resume to be reviewed.

Due the this bias issue of traditional hiring process, the practice of using NLP tools to do resume screening has become recently very popular among companies and institutions. Using NLP systems to pre-screen the resumes before reviewing the resumes could drastically reduce the amount of labor required in the hiring process and benefit both the employer and the candidates through making the hiring process fairer.

For example, a company needs 10 technicians for a construction project they carry out. However, they receives 10,000 resumes, 5,000 of which qualifies for the job. In the traditional hiring process, the company would likely review the first 50 resumes they received, select 10 out of the 50 to be hired, and then reject the rest of the candidates. With the help of automatic resume screening system, the company could first rank all 10,000 candidates by their capability, review the top 50 compatible resumes instead of the first 50 resumes that comes into sight, and then hire 10 most suitable candidates. This could help companies hire the most compatible employees, and reduce the bias in the hiring process.

Competence-Level Classification Competence-level classification could be helpful for the candidates through helping them apply to the correct level of the job, if the job has multiple levels.

For example, if a company offers junior technician, intermediate technician, and advanced technician jobs, then it is good for the candidates to use a competence-level classification system to position themselves before applying to ensure that they apply to the correct level. Such a system could also be very helpful as an auto-screening tool, as for certain job levels, company could use a competence-level classification model to screen incompetent candidates out before reviewing the resumes.

Resume & Job Description Matching Resume & Job Description matching systems could be useful for the candidates to determine whether they are a match with the job. This could ensure that they are compatible with all the requirements of the job position, maximizing their probability of getting the job. For companies, this could also be helpful in hiring people with certain skills.

For example, if the company want to hire 1 employee to connect them with their German collaborators, the company probably want to hire a person who are both proficient in the company's specialization and in German. Through a resume & job description matching system, they could easily screen candidates who, although qualifies for the job in general, does not know German. Screening such incompatible candidates could significantly speed up the hiring process.

3.3 Approach

Although the previous work [13] has achieved promising results through developing a context-aware transformer model, the data used in this work feature potential flaws. The labeled resume dataset used to develop the competence-level classification model was noisy, and the job description dataset used to develop the resume & job description model lacked variety in job descriptions, and therefore it cannot simulate the actual resume & job description matching scenario.

To address the existing issues in previous datasets, I have conducted the following in collaboration with Dr. Elaine Fisher from Emory Nursing School, and Prof. Steve Pittard from Emory Department of Biostatistics and Bioinformatics:

3.3.1 Resume Parsing and Field Concatenation

The previous work [13] used a custom-designed regular expression parsing tool to derive different sections from the resumes. They have divided each resume into 6 sections: *Profile*, *Education*, *Work Experience*, *Activities*, *Skills*, and *Others*.

In alignment with the previous work done on this dataset, we have also parsed the resume into different sections. Instead of using a regular expression, my collaborators have utilized a resume parsing tool powered by Artificial Intelligence (AI) developed by

Rchilli[1]. This tool allows us to parse the resume into more than 200 data fields, such as *Address*, *Skills*, *Certificates*, *Job Profile*, etc., leading to a more accurate parsing result.

Among all the fields parsed by Rchilli, Dr. Fisher selected 4 of the data fields (resume sections) out of the 200+ parsed resume fields based on her experience of hiring CRC for years. The 4 fields are: *Qualification* (Education), *Certification*, *Experience*, and *Job Profile*. Taking these 4 fields from the parsed resume, I concatenate these 4 fields together with the separation token `<sep>`. When I train the model, instead of using the entire resume as input, I use this extracted concatenation as the input. This could help reduce the length of the input sequence, avoiding potential information loss resulted from truncation due to the length limit of the model and helping the model focus more on the important information than the unimportant ones.

Here is a truncated example of what our parsed resume look like. The original resume from which we obtained this parsed resume and the result after concatenating these fields using `<sep>` could be found in appendix B. The resume has been anonymized through replacing the actual names with John Doe and Jane Doe.

> “Qualification”: “Master of Public Health, Behavioral Science and Health Education. Emory University, The Rollins School of Public Health, Atlanta, GA. (1998) Bachelor of Science, Human Development. University of California at Davis, Davis CA. (1992) ”,
>
> “Certification”: “”,
>
> “JobProfile”: “Research Administrative Coordinator”,
>
> “Experience”: “Emory University, School of Medicine, Division of Cardiology, Research Administrative Coordinator, Senior. Atlanta GA. (2014- present : Full-time M-F 40/hours / week) Manages administrative tasks associated with the multi-site research project Patient-Centered Approaches to Research Enrollment Decisions in Acute Cardiovascular Disease (P-CARE) . Assists in the development of the study interview guide and other study documents. Submits, gains and maintains IRB and ROC (Grady Hospital) clearance. Ensures project is administered according to research protocol. Schedules and

[1] https://www.rchilli.com/

> conducts subject interviews, focus groups, and Patient Advisory Panel meetings. Gathers and manages data. Makes and gives presentations. Serves as project liaison to multiple study sites, other departments within Emory, outside organizations, government agencies and product representatives. Assists with grants, expenditure monitoring and budgeting. Contributes to data analysis. Performs additional related responsibilities as required. Additionally, acts as Administrative Assistant for Emory Clinical Cardiovascular Research Institute (ECCRI). Processes purchase orders, requisitions, check requests, invoices, and expense reports via Compass or Emory Express for junior research faculty and cardiology research fellows. Provides calendar support. Purchases supplies for various research studies, as needed. Works closely with various departments and vendors to solve problems and facilitate timely delivery of good and services. Works with Cardiology fellows applicants during the interview process and assists with onboarding..."

3.3.2 Fixing False Labels

Error Analysis

While 73% accuracy on competence-level classification in the previous work is impressive, it could do some improvements to be deployed in actual real-world setting. Context-aware transformer model has been proven to work very well in classification tasks like this. Therefore, it is likely that problems in the training data that caused the model to work not as well as expected. To understand the data better, Dr. Fisher, Dr. Choi, Prof. Pittard, and I have conducted a sampled error analysis on the data.

First, I have requested a set of hiring guidelines for CRC job positions from the Emory Healthcare Human Resource department. This guideline clearly specifies the requirements for different levels of Clinical Research Coordinators, such as education background or clinical research experience. For details of the hiring guidelines, please make reference to the appendix A, but in short, the hiring guidelines specifies the minimal requirements for each CRC level: what degree should the candidate hold, and what type of experience should the candidate have.

Then, I randomly sampled 30 resumes from the resume dataset. For each resume, I referred to the CRC Hiring guidelines, and determine to which CRC level does the resume actually belongs. Then, based on my decision, I look back to the label and see whether it is correctly labeled according to the guidelines. For example, the resume presented in appendix B was one of the resumes that we have sampled from the resume dataset. It was originally labeled as CRC3. However, with more than 6 years of clinical research experience and a Master degree in Public Health, this resume should have been labeled as a CRC4. Such an example would be considered as a mislabeled case.

After going through all 30 resumes we have sampled, I found that 7 out of 30 sampled resumes were mislabeled. This suggests an urgency of refining the data and fixing the labels.

Revisiting Resumes

To fix the resumes, Dr. Fisher and I conducted rounds of data cleaning. Dr. Fisher, having many years of experience in hiring CRC personnels, corrected the labels in a procedure similar to how I conducted error analysis: looking into the resume, determining the correct label of the resume, and noting the correct label for the resume. After each round of Dr. Fisher's label correction, I use the refined data to train a new model, and see whether the performance of the model has improved. If the performance on the test set does not improve, then I would double check whether I have the correct version of data with Dr. Fisher.

Here is an example of how the label of a resume would be corrected. Given the resume provided in appendix B, the person would qualify as CRC4, because the person has more than 6 years of clinical research experience and a Master's degree in a public health. This satisfies the 7th requirement of hiring CRC4 according to the CRC Hiring Guidelines (make reference to appendix A for details): *MD or PhD in a scientific or health related field (Includes unlicensed US, foreign trained MD's) AND 2 year of clinical research experience*

After rounds of such cleaning, Dr. Fisher revisited 994 resumes in total, out of which 332 were relabeled with a different CRC level from their original label. Table 3.1 provides an insight to how were the labels changed in this process. The leftmost column of the table represents the original labels of what the label used to be; and the top row of the table represents to what new labels were these labels changed. For example, in the 1st column of the 4th row, the number 7 on the intersection between CRC3 and NQ suggest that in total there are 7 resumes that were previously labeled as CRC3, but after revisiting the resumes, we have changed their labels to NQ, meaning that they do not qualify for any CRC jobs.

		New Labels				
		NQ	CRC1	CRC2	CRC3	CRC4
Original Labels	NQ		36	4	2	0
	CRC1	36		60	15	0
	CRC2	9	62		17	0
	CRC3	7	29	42		1
	CRC4	0	1	0	4	

Table 3.1: Confusion matrix of corrected labels

3.3.3 Creating Job Description Dataset

The dataset should simulate the real-world scenario as close as possible, and therefore we want to have variety in our data to enable the model to handle various input text. For resume & job description matching, we need diverse pairs of resume and job description. While the previous work used a dataset of plenty distinct resumes, the work used only 4 distinct job descriptions to pair with the resumes. This lack of variety in job description made the result of the previous work on resume & job description parsing inconsistent.

To address this lack of variety, Dr. Fisher created a new dataset of 710 distinct job descriptions through reverse engineering. She first selected a number of distinct resumes. Then, she writes an unique job description that fits the job description. Taking this created job description, she find in the resume dataset 3 positive resume examples that matches this job description and 3 negative resume examples that does not match this job description. This would give us 6 resume-job description pairs for each job description we have created.

Because the job descriptions were reverse-engineered (derived from actual resumes), the dataset has a CRC level for each job description based on the CRC level to which the resume belongs. While this level is unimportant to the job description itself, since job description is considered independent once it was created out of the resume, Dr. Fisher generally picked the positive and negative resume examples from the same CRC level to which the job description, or the resume from which it was created, belongs, to enforce the model's capability of distinguishing the resumes with similar competence level.

Given the two following examples:

Resume A

> "Qualification": "Clayton State University - Morrow, GA July 2012 Bachelor of Science in Healthcare Management Georgia State University Perimeter College - Clarkston, GA May 2010 Associate of Science in Business Administration",
>
> "Certification": "CITI Biomedical Focus CITI Social and Behavioral Focus CITI Good Clinical Practices ACTSI Phlebotomy CITI ICH CITI Human Subject Research Track CITI Health Privacy and Information Security",
>
> "JobProfile": "Clinical Research Coordinator",
>
> "Experience": "Emory University School of Medicine,Atlanta, GA Clinical Research Coordinator (November 2017 - present) Identified and screened potential study participants according to study protocol Performed community outreach in order to share disease-specific information and drive recruitment and enrollment Drove enrollment of Spanish-speaking participants by utilizing Spanish fluency Maintained daily, weekly, and monthly inventory Worked closely with investigator team to meet recruitment goals and ensure data quality Performed phlebotomy according to study protocol Processed and shipped study specimens according to good clinical laboratory practice and Environmental Health and Safety Office (EHSO) guidelines Conducted and collaborated with Neuropsychologist on cognitive testing for study participants Performed the following tasks related to study visit : consenting, vital signs, gait analysis, and retinal imaging Assisted with microbiome and lumbar punctures Conducted

> follow up phone calls after study visits Facilitated the hiring process of new employees by reviewing resumes, conducting interviews, and providing feedback to hiring manager Assisted with new employee orientation / training, Atlanta, GA Administrative Assistant / Patient Care Coordinator (September 2013 - November 2017) Maintained 100% compliance by departing inpatient and outpatient correspondence in a timely fashion Directly provided administrative support to 2 surgeons who rank in the top 10% patient satisfaction surveys..."

Resume B

> "Qualification": "PhD in Chemistry, University of Cape Town, South Africa 2013-2018 Thesis title: Repositioning fusidic acid for tuberculosis: semi-synthesis of analogues and impact of mycobacterial biotransformation on antibiotic activity. Combined knowledge of chemistry and biology in the repositioning of an antibiotic as an antituberculosis agent.",
>
> "Certification": "",
>
> "JobProfile": "Postdoctoral research fellow",
>
> "Experience": "Postdoctoral research fellow, University of Cape Town, South Africa July 2017-present Applied lipidomics and proteomics in the validation of novel drug targets for Tuberculosis drug discovery at the Drug Discovery and Development Centre (H3D) . Skills learnt include : Comparative lipidomics and proteomics, molecular biology and microbiology. Visiting research fellow, University of Leeds, UK October-December 2018 Expressed, purified and characterised EfpA from Mycobacterium tuberculosis using E.coli as a host expression system. Next Generation Scientist fellow, Novartis, Basel, Switzerland June-August 2015 Expressed, purified and characterised a human protein using E.coli as a host expression system and conducted a fluorescence-based biochemical assay for anticancer drug target validation. Intern, Consortium for National Health Research, Nairobi, Kenya January-July 2010 Isolated, purified and characterized organic extracts from medicinal plants for antimalarial drug discovery. Skills learnt include chromatographic separation and spectroscopy Tutor and Demonstrator, Department of Chemistry, University of Cape Town 2011-2017 Tutored chemistry and facilitated practical sessions to undergraduate students from the faculties of engineering and health sciences. Pharmacist, Ministry of Medical Services, Kenya 2009-2011 Supervised staff within the pharmacy department and ensured proper dispensing practice. Managed medical supplies within six Kisumu West district healthcare facilities. Provided advice to clinicians on patient management..."

In the previous work, these two resumes, along with hundreds of other resumes being labeled as CRC4 in the labeled resume dataset, would all be associated with the generic CRC4 job description, which is the following:

> Recruits, orients, and supervises research administration staff or independently manages the most complex research administration activities associated with the conduct of clinical trials. Manages a large or multiple smaller clinical research projects. Manages clinical trials related information systems. Supervises the implementation of and adherence to study protocols. Monitors expenditures and adherence to study budgets and resolves CAS issues. Educates research staff on established policies, processes and procedures. Determines effective strategies for promoting/recruiting research participants and retaining participants in long term clinical trials. Periodically audits operations including laboratory procedures to ensure compliance with applicable regulations; provides leadership in identifying and implementing corrective actions/processes. Plans, identifies, and handles study related equipment and facilities needs. Provides leadership and expertise in identifying and completing research grants, study materials, brochures and correspondence. Develops and submits grant proposals. Leads or chairs committees or task forces to address and resolve significant issues. Performs related approved responsibilities as required.

However, now that Dr. Fisher developed this job description:

> Participate in or lead day to day operations of clinical research studies, perform a variety of duties involved in the collection, compilation, documentation, and analysis of clinical research data. May oversee the work of junior staff and train or mentor others in clinical research tasks. A minimum of 2 years clinical research experience, hold a bachelors or higher academic degree. Experience in memory and aging research.

Resume A matches this specific job description because the person had experience in memory and aging research, whereas resume B does not match this job description because the person had no experience in memory and aging research. Although both resumes belong to CRC4, in which case they would have both been considered as a match in the previous work [13], we have now added this degree of variety into this dataset. This variety forces the model to learn to extract important details, such as experience in memory and aging research, from the job description, and use that to decide whether the resume matches the job description, which would not have been the case in the previous work.

3.3.4 Encoding Resumes with Labels

Competence-Level Classification

For competence-level classification, each resume is binded with an annotated label of CRC level. Therefore, when training the model, I could directly formulate this task as a simple linear-directed classification problem, where I use the field concatenation, as mentioned above, as the input of the model, and use the label as the output.

Instead of extracting the desired fields from the parsed resume every time and concatenating them, I chose to store the desired fields with the concatenation and the label, and use that as the data for competence-level classification.

Here is a representation of a encoded labeled resume, where `CONTENT` is the placeholder for the actual field content in the parsed resume:

```
"Qualification": "CONTENT",
"Certification": "CONTENT",
"Experience": "CONTENT",
"JobProfile": "CONTENT",
"Competence-Level Label": "CRCX"
```

Resume & Job Description Matching

At first, Dr. Choi and I have attempted to assign a job description and a match label to each resume. For example, given the example with the placeholders above, if this particular resume also appears as a positive or a negative example in the newly developed job description dataset, then we append the job description and the match label to what I have encoded for this resume for competence-level classification. Given the placeholding example provided above for competence-level classification, this resume would now be stored in the following json format:

```
"Qualification": "CONTENT",
"Certification": "CONTENT",
"Experience": "CONTENT",
"JobProfile": "CONTENT",
"Level_Label": "CRCX",
"JobDescription": "CONTENT",
"Match_Label": "YES/NO"
```

However, such encoding did not work, because different from competence-level labels, the job descriptions are not 1-to-1 binded with the resume. In other words, a resume could appear in multiple job descriptions as either positive or negative examples. For example, the resume B we have provided in the previous section is a negative example for the job description that we have newly created. However, this particular resume is a positive example for this job description:

> Lead day to day operations of complex and multiple study teaMS degree conducting clinical research. Perform a variety of complex duties involved in the collection, compilation, documentation, and analysis of clinical research data. Lead others in navigating the clinical research environment. Leads or participates in a variety of unit, department, or division-level initiatives. Oversee the work of CRCs and other research staff. MS degree and a minimum of 3 years of clinical research experience. Certification from a professional clinical research organization is required on hire. Preferred qualifications: serves as an expert resource to teaMS degree, experienced creating and managing large data-based studies, experience with phase I-IV clinical drug trials, compliance and regulatory monitoring and reporting, overseeing quality assurance initiatives, and budget development and management.

If I encode the job descriptions and the match labels like the competence-level labels on a resume-by-resume basis, a great data loss will be resulted from the encoding, because a resume would be only referred to one of the job descriptions in which it appears as a positive or a negative example.

To address this problem, I have encoded the data for competence-level classification and that for resume & job description matching separately. In a set of data dedicated for competence-level classification, I encoded the resumes as exactly described in the previous subsection.

For the resume & job description matching, I encoded it on a description-by-description basis, meaning that each resume could appear in the data multiple times with no upper limit as long as it appears as a positive or a negative example in the job description dataset, but each job description would appear for at most 6 times, in 3 positive examples and 3 negative examples. A toy example would be

```
"Qualification": "CONTENT",
"Certification": "CONTENT",
"Experience": "CONTENT",
"JobProfile": "CONTENT",
"JobDescription": "CONTENT",
"Match": "YES/NO"
```

3.4 Experiments

3.4.1 Data Split

Labeled Resume Dataset

As mentioned in the previous section, the model for both tasks were developed on a dataset of 3425 resumes that were split into 3 sets: training set, validation set, and test set. In addition to the data used by the previous work, I have received another 1500 of labeled resumes from Emory HR department. In order to align our result with the previous work, we have added this entire new 1500 labeled resumes into the training set, which is used to later train the model. There are 2 advantages of doing this: (a) with more data used to train the model, the model could potentially achieve better performance; and (b) since we add the entire new 1500 resumes into the training set, the validation set used to tune the model and the test set used to test the model remains unchanged. This would allow us to directly compare the model with the ones in the previous work. Table 3.2 gives us a comparison between our data split and the split of the previous work.

	TRN	DEV	TST	Total
Previous Work [13]	2565	344	516	3425
Our Work	4065	344	516	4925

Table 3.2: Comparison of Data Splits

After I add the additional 1500 resumes into the training set, table 3.3 provides a detailed elaboration on the resumes for each CRC job level. Apart from the 4 CRC levels, there are also resumes that should be classified as NQ, suggesting that the resume does not qualify for any of the CRC job level.

	TRN	DEV	TST	**Total**
NQ	648	48	72	768
CRCI	2203	202	302	2707
CRCII	511	38	58	607
CRCIII	658	52	79	789
CRCIV	45	3	5	53

Table 3.3: data distribution of our dataset

Job Description Dataset

The job description dataset is not comparable between our work and the previous work, because the previous work contained only 4 distinct job descriptions, while our contains 710. The benefits of such a dataset with variety in job descriptions has already been elaborated in previous sections.

As mentioned in the previous section, for each job description, we have found 3 positive examples and 3 negative examples from the resume dataset. Because we try to stick to the CRC level to which the job description belongs, there are cases where we cannot find 6 such examples, especially for job descriptions created from resumes that are labeled CRC3 or CRC4 (since less resumes are labeled as CRC3 or CRC4). Therefore, the total number of resume-job description pairs is 4112 instead of 4260. The details of the splits of resume and job description pairs was conducted based on a 80-10-10 ratio, which was elaborated in table 3.4.

	TRN	DEV	TST	**Total**
Y	1654	196	208	2058
N	1645	199	210	2054
Total	3299	395	418	4112

Table 3.4: Job Description Data Split

3.4.2 Modeling

Using the datasets for Competence-Level Classification and Resume & Job Description, we have developed BERT [5] and RoBERTa [15] models. The reason we chose these models is that they were the state-of-the-art models of doing sequence classfication by the time when we conducted this project. The Huggingface [2] package is used to access the `bert-large-cased` model and `roberta-large` model. The models are trained using a single Nvidia A6000 GPU. The number of epochs between 1-10 and the learning rate between 1e-2 to 1e-5 were attempted during hyperparameter tuning. We have found that a RoBERTa model trained 9 epochs at the learning rate of 2e-5 works the best for Competence-Level Classification, and the BERT model trained 3 epochs at the learning rate of 1e-5 works the best for Resume & Job Description Matching.

Pre-Training Using Resumes

The two transformer model structures we have attempted, BERT and RoBERTa, are all trained through a 2-stage process: pre-training and fine-tuning. During the pre-training stage, the model is trained on a large corpus of various domain text data

[2] https://huggingface.co/

through unsupervised approaches like the Masked Language Modeling (MLM) [5]. And then in the fine-tuning step, the model learns the specific task that we hope to deliver through the model through supervised learning.

In our case, when we train the competence-level classification model and the resume & job description matching model with the competence-level labels and the match labels, we are fine-tuning the model through supervised learning with these labels as supervision.

For our project, in addition to fine-tuning BERT and RoBERTa model, we have also attempted to pre-train the model using all of the resumes we have in the dataset to first pre-train the model. To do this, we first pull the pre-trained BERT or RoBERTa model with a large corpus (typically Wikipedia corpus) from Huggingface, and then we feed the model the unlabeled resumes we have parsed and concatenated. After we feed the unlabeled resumes as a "second pre-training step," we fine-tune the model using the labeled data. Technically, this should help the model better understand the structure of a resume before it is asked to do classification.

Staged Learning

We have attempted staged learning in doing competence-level classification. In our concatenated input of the 4 fields, we have separated these fields with the separation token `<sep>`, and we would append the class token `<CLS>` at the beginning of our concatenation. When the BERT model or the RoBERTa model generates output embeddings for our concatenated sequence, the model would generate the embeddings for each input token and these special tokens as well.

When we trained the model, we instruct the model with the input and the correct output. In the expected correct output, we make the competence-level label the expected output of the `<CLS>` token, as shown in figure 3.1.

And when we use the model to predict the competence-level of a resume, we input

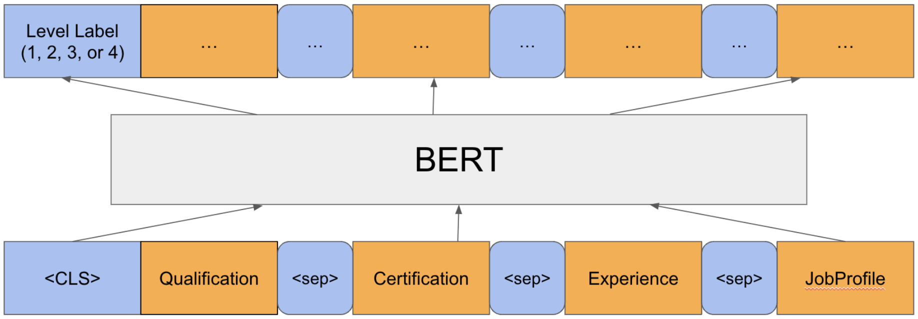


Figure 3.1: Representation of training an end-to-end BERT classification model

the concatenation into BERT, and pull out the embedding generated by BERT on the `<CLS>` token. This embedding would be our final prediction for the model, because the model was trained in that way, as shown in figure 3.2

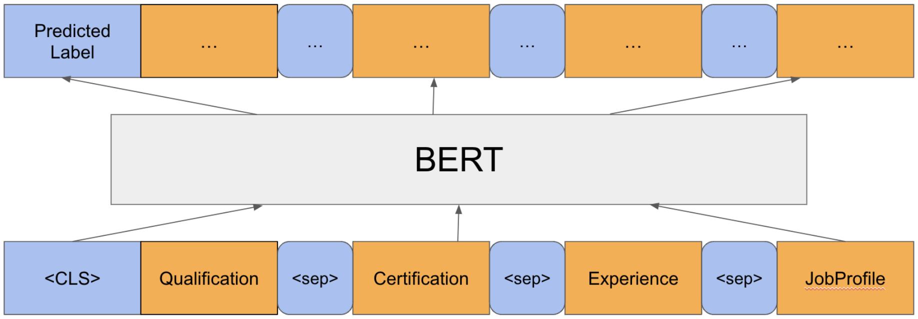


Figure 3.2: Representation of using an end-to-end BERT classification model to predict the competence level of a resume

By staged learning, we mean to separate the concatenated input as different stages. When we pull the generated embeddings from the model, instead of pulling the embedding of the `<CLS>` token, we pull the embedding of the `<sep>` tokens. And we apply a stage learner on top of the embeddings from the `<sep>` token to generate the final classification result. Figure 3.3 provides a comprehensive elaboration of the staged learning structure.

I have designed a simple stage learner, where I do a weighted sum of the 4 embeddings from the 4 `<sep>` tokens. During training, I backtrack the weights first through the stage learner, and then through the BERT model as one would train a

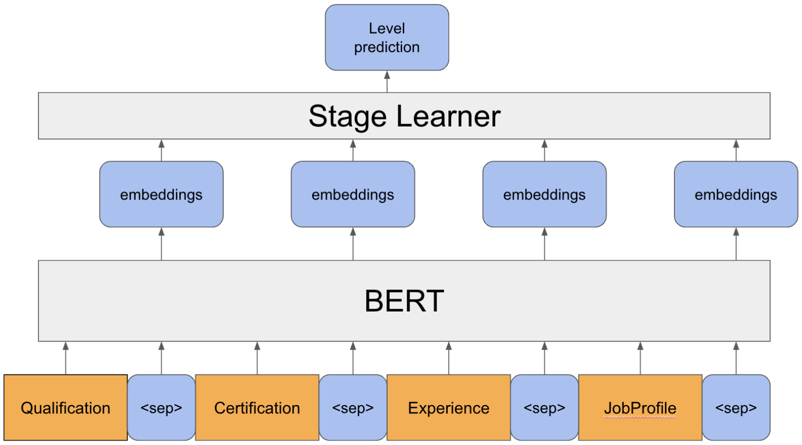


Figure 3.3: Representation of a staged learning structure on top of BERT

normal BERT model. The structure of the staged learner is represented in figure 3.4.

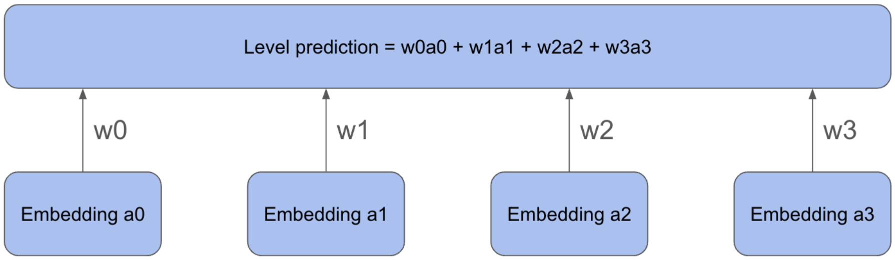


Figure 3.4: Staged learner structure

Nevertheless, the result was not as good as the result without the staged learner. Furthermore, because I have just started doing research by that time, I do not have these results as I am writing my thesis. But trust me, it did not work. It could well be that a better staged learner structure could help enhance the performance.

3.4.3 Results

Competence-Level Classification

Data cleaning for the labeled-resume dataset of CRC levels resulted in a significant improvement in model performance. Whereas before the data cleaning, the baseline model, which is a BERT model, could only achieve 69% accuracy, after cleaning up the data and fixing the labels, we were able to achieve 74%. Furthermore, out naive-RoBERTa model has a significant advantage over even the context-aware model developed in the previous work, which was the best model discussed. This shows that data cleaning is very effective in enhancing model performance. Although the improvement is not drastic, note that we have only used a simple BERT-based model. This model is much less complicated and smaller than the context-aware model, making it easier to train and deploy.

	ACC
Previous-BERT	69.06 (± 1.56)
Previous-Context-Aware	73.26 (± 0.16)
Ours-BERT	74.83 (± 0.53)
Ours-RoBERTa	75.65 (± 0.92)

Table 3.5: Model Accuracies for T1

Resume & Job Description Matching

For resume & job description matching, the results between our work and the previous work is not comparable. Because the variety of job description is introduced in our task, the data is drastically different, and our model is facing a significantly more difficult dataset than the previous work. While the model trained in the previous work would only need to learn the traits of the 4 job descriptions they have used, in our case the model would actually have to learn to understand the relation between the resume and the job description, because there are now 710 distinct job descriptions, and thus there is no way to memorize all of the job descriptions given the size of our model.

Our model has achieved roughly 73% accuracy on our newly developed job description dataset. We ran this experiment for 5 times to offset the randomness in model parameter initialization. Table 3.6 elaborates the result of each attempt.

Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5	Average
77.38%	73.78%	73.78%	71.98%	71.98%	73.78%

Table 3.6: Resume & job description matching model performance results

3.4.4 Code Packaging

Because the project is an application-oriented project and our goal is to deploy our model, My thesis advisor and I packaged our code into `eclair-transformer`[3]. To use the code, encode the labeled resume dataset and the job description dataset according to the documents of the package, which could be found either in the `readme` file upon clicking the url in the footnote or in the previous subsections where we have discussed how we encoded the json file for the two tasks.

3.5 Analysis

3.5.1 Confusion Matrix

For competence-level classification, we have derived a confusion matrix of how the model performs. Table 3.7 present details of the model output on the test split of our cleaned dataset. The leftmost column are what the gold true labels are the resumes, and the top row represents what the model predicts the resume to be. For example, the number 35 on the top left corner represents that 35 resumes that are labeled as `NQ` has been predicted correctly; and the number 29 next to it represents that 29 resumes that should actually be `NQ` has been wrongly predicted as `CRC1` by the model.

3.5.2 Ablation Studies

During the experiments, there are two steps that could influence model performance on competence-level classification: the newly-added 1500 resumes into the dataset,

[3] All our codes for training and predicting are publicly available at `https://github.com/emorynlp/eclair-transformer`

		Predicted Labels				
		NQ	CRC1	CRC2	CRC3	CRC4
Actual Labels	NQ	35	29	1	4	0
	CRC1	10	270	13	17	0
	CRC2	0	16	21	14	0
	CRC3	0	14	10	58	0
	CRC4	0	0	0	4	0

Table 3.7: Confusion matrix of the model output on the test set

and the labels corrected by Dr. Fisher. Compared with the previous work, we have also attempted to pre-train the model before fine-tuning the model with the labels. Therefore, there is a need to identify which of the 3 factors contributed to the improvement in model performance.

In order to understand the contribution of each factor to model performance, we have conducted an ablation study. All of the experiments are ran on our refined data after Dr. Fisher corrected the labels. However, we have tried to take off the newly added 1500 from the dataset, and we have tried to train the model with and without pre-training the model with resumes. The results are summarized in the table 3.8.

	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5	Average
BERT w/o 1500	72.43%	73.59%	73.01%	75.73%	75.92%	74.14%
BERT w/ 1500	73.59%	74.37%	75.53%	74.56%	76.12%	74.83%
RoBERTa w/o 1500	74.17%	74.95%	74.76%	73.59%	74.37%	74.37%
RoBERTa w/ 1500	74.37%	75.73%	76.89%	75.34%	75.92%	75.65%
BERT-pretrained w/o 1500	73.01%	72.43%	73.01%	71.46%	72.43%	72.47%
BERT-pretrained w/ 1500	75.53%	74.95%	74.37%	73.79%	73.01%	74.33%
RoBERTa-pretrained w/o 1500	74.17%	71.46%	72.82%	73.20%	74.76%	73.28%
RoBERTa-pretrained w/ 1500	75.53%	75.15%	73.79%	74.56%	73.20%	74.45%

Table 3.8: Impact of the added 1500 resumes and second pre-training with resumes

The 5 training attempts are all exactly the same (same training data and same test set). The model performance slightly fluctuates between attempts because the model parameters were randomly initialized, and therefore some training attempts might result in slightly better performances than others.

From the results we can see that adding the 1500 resumes into the dataset contributes only slightly to the increase in model performance (about 1%). For example, for un-pretrained BERT model, taking out the 1500 resumes that were added to the training set, the model still performs somewhere around 74%, which is still significantly higher than 69% in the previous work before the wrong labels were corrected. Pre-training the model with resume data actually resulted in worse model performance. Therefore, the results we have reported in the previous subsection were those without pre-training the model with unlabeled resume data.

3.5.3 Result Human Analysis

After training the model, I have conducted another round of error analysis on the labeled resume dataset that was used to train the competence-level classification model. For this round, instead of randomly selecting resumes from the dataset, we have focused on the resumes on which the model made the wrong predictions. These resumes are generally either ambiguous or harder.

Out of 516 labeled resumes in the test set, the model correctly predicted 384 of the resumes, and wrongly predicted 132 resumes. Through looking into these resumes in the same way how we conducted error analysis before training the model, as mentioned in the Approach section, we found that 18 resumes out of 132 falsely-predicted resumes have wrong labels. Compared to 7 out of 30 before data cleaning, the data quality has significantly improved through data cleaning. This improvement in data quality strongly positively correlates with the model performance.

3.5.4 Limitation and Future Works

Modeling

By the time when we conducted this project, BERT was the state-of-the-art model for doing sequence classification. Recently, the advancements of LLM suggested promising results in various sequence-to-sequence tasks, and a recent study has shown that sequence-to-sequence model could do classification as well as they output sequences [7]. This suggest a potential that LLMs could potentially do better on competence-level classification and resume & job description matching with the right problem formulation.

Staged Learning

When I attempted the staged learning experiments, I was just starting to do research in computer science. Therefore, I have only attempted a very simple staged learner, and I did not record the results of my staged learning experiments. Studies have shown that staged learning could enhance model performances in classification tasks [14], and therefore more experiments on applying staged learning on our tasks could potentially improve model performance.

Chapter 4

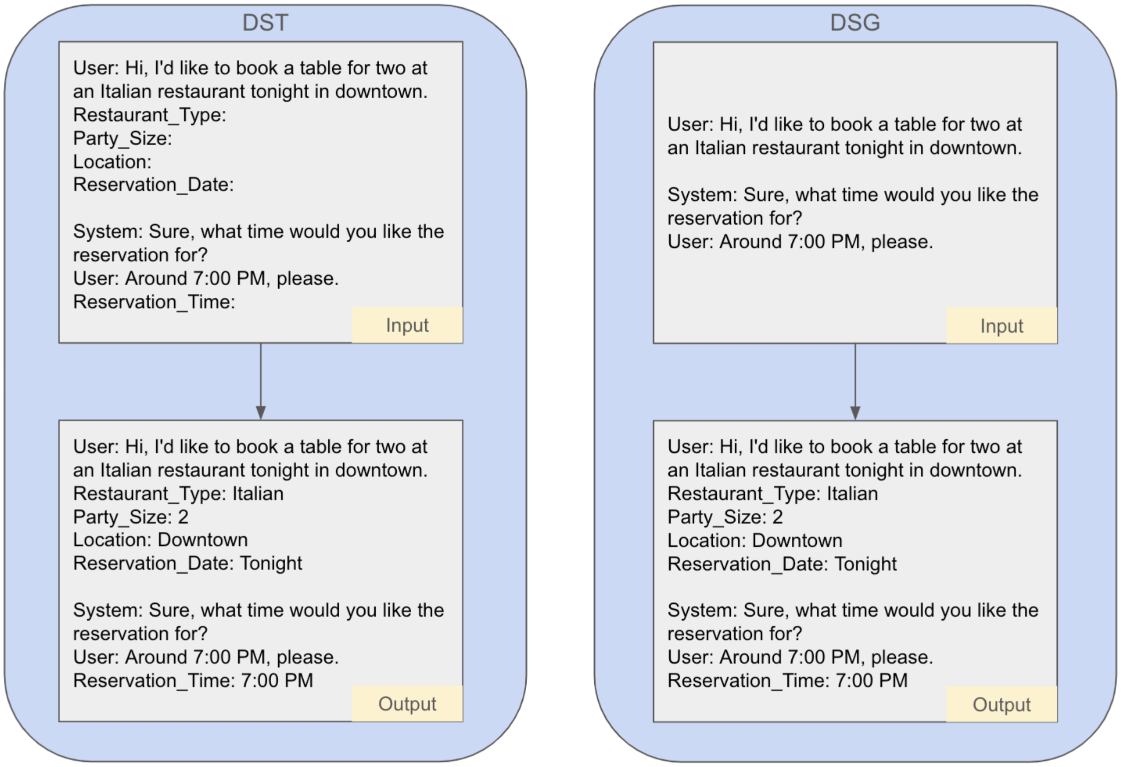
Dialogue State Generation

4.1 Task Definition

To address the shortcomings of Slot Discovery, in this project, we propose Dialogue State Generation (DSG), which is to use a model to automatically do DST without given slot names: the model decides what information are important from the input, and automatically generate slot-value pairs based on the input.

This task addresses the shortcomings of the SRL. Using a pre-trained DSG system, we would not need any dialogue data in the context where we want to deploy the chat bot. The DSG system would generate the slot names and values in such a zero-shot scenario. Also, because it is slot-value pairs that are being directly generated from the input, we would not need to figure out the slot names based on the clusters of values.

Figure 4.1 elaborates the difference between DST and DSG through an example. Basically, DSG is to extract information from the dialogue data without telling the model what to extract in the input prompt.



User: Hi, I'd like to book a table for two at an Italian restaurant tonight in downtown.
Restaurant_Type:
Party_Size:
Location:
Reservation_Date:

System: Sure, what time would you like the reservation for?
User: Around 7:00 PM, please.
Reservation_Time:

User: Hi, I'd like to book a table for two at an Italian restaurant tonight in downtown.
Restaurant_Type: Italian
Party_Size: 2
Location: Downtown
Reservation_Date: Tonight

System: Sure, what time would you like the reservation for?
User: Around 7:00 PM, please.
Reservation_Time: 7:00 PM

User: Hi, I'd like to book a table for two at an Italian restaurant tonight in downtown.

System: Sure, what time would you like the reservation for?
User: Around 7:00 PM, please.

User: Hi, I'd like to book a table for two at an Italian restaurant tonight in downtown.
Restaurant_Type: Italian
Party_Size: 2
Location: Downtown
Reservation_Date: Tonight

System: Sure, what time would you like the reservation for?
User: Around 7:00 PM, please.
Reservation_Time: 7:00 PM

Figure 4.1: Difference between DST and DSG input prompts

4.2 Task Motivation

4.2.1 Shaping Chat Bot Behavior

Conversational Artificial Intelligence (AI) is a very popular piece of technology that is very close to people's daily life. From Siri and Alexa to ChatGPT, conversational AI is more and more facilitating people in life. Controlling the behavior of conversational AI systems, or chat bots, as many may call these systems, is very important for both practical and ethical reasons. Practically, we want to control the behavior of the bot so that it could be more helpful and useful in the applications to which we deploy the bot. Ethically, we want to prevent the bot from spitting out unethical things that could potentially cause harm to people and the society.

Traditionally, chat bots are built based on hard-coded state machines, meaning that the chat bot would only respond properly to the input that it has been programmed to response. For example, if I want a chat bot based on a state machine to react to people's "hello"s, I will have to program the chat bot: when someone says "hello" to you, say "hi." Once it is programmed, the chat bot would respond "hi" to every single "hello" it receives.

Recently, the highly-intelligent Large Language Model (LLM) changed the world of chat bot development. An LLM is a type of artificial intelligence system designed to understand and generate human-like text. These models are trained on vast amounts of textual data and are capable of performing a variety of natural language processing tasks, such as text generation, translation, summarization, question answering, and more. LLMs are characterized by their complexity, often consisting of billions of parameters, which enables them to capture intricate patterns in language and produce coherent and contextually relevant text.

Chat bots based on LLM are easier to develop and have better performances compared to traditional state machine chat bots. However, because of the size of LLM and the nature of it as a deep learning model, we do not have full control over its generated content. Unlike the chat bots based on state machines, where every state of input is hard coded, LLMs generate the response independently, and therefore is harder to control.

4.2.2 Dialogue State Tracking

Seeking to control the behavior of deep-learning-based sequence-to-sequence models like LLM, NLP researchers have proposed Dialogue State Tracking (DST), where the model first tracks the state of the input, and use the tracked state to develop the prompt for LLM to have it behave correspondingly. DST by its essence is to extract useful information from the dialogue based on the given slot names. For example, if one is to develop a chat bot for a restaurant that automatically makes reservations for the customers calling into the restaurant, one could decide that "reservation time" and “party size” are required for booking a table. With knowing what information we want, one could make up a prompt like

```
INPUT TEXT.
party size:
reservation time:
```

One could then use this prompt to prompt an LLM, and then have the LLM automatically output the extracted information from the customer’s call.

For example, if the customer called and says “Hi, may I make a reservation for 4 people? We will arrive at around 6:30.” Then the prompt will become

```
Hi, may I make a reservation for 4 people? We will arrive at around 6:30.
party size:
reservation time:
```

And we could expect the LLM to, for example, fill out the blanks and output something like:

```
Hi, may I make a reservation for 4 people? We will arrive at around 6:30.
party size: 4
reservation time: 6:30
```

Research have shown that few-shot DST works better than zero-shot DST. That is to say, if in the prompt we add in a few examples, the LLM is likely to perform better. In the example above, instead of prompting the model like what we did above, we could prompt the model through adding in an example like

```
Here is an example of what I want you to do:
I want a table for 5 people to be ready for me at 5pm.
party size: 5
reservation time: 5pm
<sep>
Based on the example, extract the information from the following:
INPUT TEXT.
party size:
reservation time:
```

With such a prompt with an example, the model is likely to perform better than given the prompt with no example.

Through doing DST, we are giving the LLM a direction of what we want instead of having it generate whatever it thinks is the most important. Through giving it what we want, we made the LLM more useful to our downstream application.

While DST is good for tailoring the LLM according to the application scenario, it has a couple major drawbacks that makes it not very convenient. Doing DST requires us to understand what information we want in advance (in the example above, we must know that we want to extract "reservation time" and "party size" before we develop the prompt); prompt engineering the LLM could also be very challenging, as the behavior of LLM is, up to today, still not fully interpretable (the difference in one or two words in the prompt could completely change the behavior of LLM).

4.2.3 Slot Discovery

To address the issue of DST must have the information to be extracted in mind in advance, NLP researchers have proposed Slot Discovery (SD), where slots could be automatically discovered from existing dialogue data instead of figuring the slots out by human. In the same example we have discussed above of a restaurant reservation chat bot, with the help of slot discovery, we would not have to figure out that "reservation time" and "party size" is what we want to extract from each phone call. Instead, we could put, for instance, the transcriptions of 300 recorded phone calls into Semantic Role Labeler (SRL), and extract words from the dialogue that could potentially be useful for the task that we want to do. Then, through clustering the values, we could potentially see what information we want based on the results discovered by SRL. For example, in developing the restaurant reservation chat bot, if the combination "`number` pm" (`1` pm, `2` pm, etc.) frequently appear in the 300 recorded phone call transcriptions, then through looking at the clustering results, we could see that there are so many people mentioning the time, and therefore conclude that we probably want to extract "reservation time" from the dialogue.

Slot discovery is very useful when we do not have an idea of what information we want to extract in the first place. However, slot discovery requires a large amount of dialogue data to extract potential clusters, which we do not always have when developing a chat bot. And after it obtains the clusters, we still have to figure out the slot names (like "reservation time") by ourselves.

4.3 Approach

4.3.1 DSG Zero-shot with GPT Pipeline

Mr. James Finch conducted this part of our research. If zero-shot GPT could do DSG, there is no need to develop any other DSG models that requires further pretraining. To prompt engineer GPT for the best result, Mr. Finch has divided the task into two parts, through which GPT could perform better than the one-step approach. He first ask GPT to break down all the valuable information into Question-Answer pairs, and then instruct GPT to translate the QA-pairs that it developed in the last step into slot-value pairs, which are the output for the task DSG. This approach serves as a baseline DSG model, with which we would later compare our model.

4.3.2 Developing New Dataset for Dialogue State Generation

Existing datasets are not suitable for training DSG models. Those existing datasets that has lots of information, such as the Schema Guided Dialogues (SGD) [23] and the MultiWOZ [3], typically do not cover many domains. In DSG, we hope that the model could recognize information of any domain. This lack of variety in domains makes these datasets unsuitable for training DSG models. Other existing datasets that features a wide variety of domains, such as the Blended Skills Talk [24], typically they suffer from a lack of information. Therefore, there is a good reason to develop a dataset that specifically works well for DSG, spanning lots of domains and contains lots of information simultaneously.

Traditionally, creating these datasets would be very expensive: it might not worth the cost to develop such a dataset specifically for DSG. However, recently LLMs have been proven to be very good at generating such data [17]. With this low-cost option at hand, we assume that we could develop a more robust DSG model at a reasonable cost.

Mr. Finch led the development of the data generation pipeline using GPT. He designed a 3-step approach, where he first instruct GPT to generate a set of scenario description, each of which describes a scenario like `Parent talks to teacher about afterschool programs`. Then, each scenario description is provided to GPT to generate a comprehensive list of information about the scenario, because the dialogues that GPT directly generated from the scenario descriptions are lack in valuable information. Finally, providing both the scenario description and the information list GPT has generated based on the description, Mr. Finch instructs GPT to generate dialogue data along with slot-value pairs as dialogue states, which is the final dialogue data generated by GPT.

4.3.3 Analysis on GPT-Generated Data

After generating the data, I have conducted a brief error analysis on the GPT-generated DSG data. While this part is not as significant as the previous parts, I am singling out this part as a stand-alone subsection because I did this part and this is my honor thesis.

We have randomly sampled 2 dialogues from the generated data with the first dialogue featuring 38 turns and the second dialogue featuring 28 turns. For each turn of the dialogue, there exists on average 7-10 slot-value pairs. For each slot-value pair in the data, I follow a guideline created by Mr. Finch on errors to be checked, and I notate them in the file if the generation is not good. The guideline could be found in appendix C.

Table 4.1 provides an example of how the GPT-generated data was presented to me when I did the error analysis and how I notated the potential fixes in the file. Here, the first 3 columns are hallucinated information that does not exist in the focal turn. Therefore, I have notated them, according to the baseline, with double pipes (||), fixing them to `Not mentioned`.

Excellent. Now, let's move onto brand identity and messaging. Sally, could you give us some ideas on the messaging we should use for each product?			
	total budget	$500, 000 || Not mentioned	The total budget allocated for both products combined.
	smartwatch allotment	$300, 000 || Not mentioned	The amount of budget allocated specifically for the smartwatch campaign.
	protein allotment	$200, 000 || Not mentioned	The amount of budget allocated specifically for the protein shake powder campaign.
	messaging ideas	?	Different ideas for the messaging that can be used for each product.
	brand mission		The tone and style of communication.
	target market behavior		The behavior of the target market for the new line of smartphones, such as being constantly busy and heavily dependent on smartphones.
	smartwatch target audience		The primary target audience for our smartwatch is individuals, particularly males and females aged 25-40, who lead an active lifestyle and are interested in fitness tracking.
	advertising team understanding		The advertising team understood that the communication should have a high-end luxury feel.
	specific direction		The Product Brand Manager has a specific direction in mind for the advertising themes and concepts, such as focusing on powerful possibilities, innovative features, or cutting-edge technology.

Table 4.1: Example of a turn in the GPT-generated data

When the data was created, it was designed for training both DST and DSG models. Therefore, there are slot-value pairs where intentionally the slot is not in the turn and the value is empty. The reason such slot-value pairs exist in our generated dataset is to educate DST models not to generate any dialogue state for slot names that are not present in the focal turn. When we train DSG model on our dataset, we could always easily remove these slot-value pairs with non-existent slot names and empty values, so that the DSG model would only be trained to generate useful information. In the error analysis, such empty-valued slot-value pairs are also examined. If the value is empty at where it should exist some information, I will also notate it with double pipes.

Our data analysis has proved the data to be of good quality. The most prevalent issue is state leakage, which will be discussed in more details later. In short, when we do DST or DSG, we want the model to generate only the information updates from the focal turn. We don't want the model to extract information that are not in the present turn. State leakage is when the model outputs information from last turns. In the example provided in table 4.1, the top 3 slot-value pairs are actually results from leakage from the previous turn:

> Sure thing. We have a total budget of $500,000 for both products combined. We're allotting $300,000 for the smartwatch campaign and $200,000 for the protein shake powder campaign.

We actually do not want this to happen. Out of 183 slot-value pairs with non-empty value in the 2 dialogues we have sampled, 29 slot-value pairs were notated as potential fixes, 19 of which are results of state leakage. While this is not a large portion of the data (roughly 10%), we could see how this type of error gets passed on to the model trained on this dataset in our result analysis on model output. Nevertheless, considering the portion of errors in the dataset, we consider the data quality generally good. It would be capable for training DSG models.

4.4 Experiments

4.4.1 Dataset

The DSG dataset Mr. Finch generated using GPT is used to train an end-to-end DSG model, where DSG is framed as a sequence-to-sequence task. The input sequences are the dialogue data that we have generated; while the output sequences for each turn are the slot-value pairs that the model generated from the focal dialogue turn.

In comparison with the dataset that we have developed specifically for DSG, we have also trained the same model on the Schema Guided Dialogues (SGD) dataset as a baseline model.

For both of our newly-developed DSG dataset and the SGD dataset, Mr. Finch have conducted a hold-out by domains to create the evaluation set out of the dataset. This means that the model would be trained on dialogues that features only information from domains other than the hold-out domains, and then would be tested on dialogues that features at least information from one of the hold-out domains. This would ensure that when the model is generating dialogue states from dialogues from the evaluation set, the model is guaranteed to handle untrained domains and information.

Mr. Finch has also attempted the popular MultiWOZ [3] dataset. However, the result was very unsatisfactory based on my brief pilot analysis after training the model on this dataset, so we excluded it from our work.

4.4.2 Modeling

After separating the evaluation set from the training set using domain hold-out technique, Mr. Finch used the training set to train a `t5-3b` model for both datasets, where the Huggingface package was again used to access the `t5-3b` base model.

4.4.3 Results and Evaluation

Human Evaluation Motivation

For DSG, analyzing the slot-value pairs as a linearized sequence, as one would do for DST, would not work, because it fails to account the slot-value structure. Furthermore, because the states are automatically generated by the model instead of given by the data, the results of comparing them as linear sequences would not be comparable with DST models.

Different from competence-level classification and resume & job description matching, where the former is a multi-label classification problem and the latter is a binary classification problem, DSG is a sequence-to-sequence task, meaning that while the input is text, the output is also text. This makes evaluation of the model more difficult than classification problems, because for classification problem there is a set of gold labels. If the model generates the correct label, it is good; if the model generates the wrong label, it is bad. However, for sequence to sequence model, such an universal standard do not typically exist. Therefore, for sequence-to-sequence tasks, error analysis is more than an overview of the data quality. It is more of an important part of the evaluation metric, working like accuracies or F1-scores for the model.

Human Pilot Analysis

To quantify the performance of the model, I have created numerous error categories, each of which describes a type of error that a slot-value pair features. If a slot-value pair does not fall into any of these error categories, we consider the pair as good quality. The error categories are not mutually-exclusive, meaning that one slot-value pair could potentially fall into multiple error categories. However, I count a slot-value pair as good only if the pair does not fall into any of the error categories.

The error categories are: `Hallucination`, `Leakage`, `Inaccurate`, `Redundant`, Partial, Imprecise, and Missing. The definition, example, and the performance of each model are elaborated in table 4.2. The SGD-DSG represents the DSG model trained on the SGD dataset; the GPT-Pipe represents the capability of the GPT pipeline that we developed to create the data; and the E2E-DSG is the model trained on our GPT-generated DSG dataset.

Mr. Finch randomly sampled 100 turns from every model's output on the evaluation set, and then I looked into these turns and generations. Note that when I examined the output associated to a certain turn, I would only look at the focal turn and the slot names. Then, I examined whether each slot-value pair falls into any of the categories we have proposed. If the slot-value pair is a good piece of information that does not fall into any of the error categories, I mark them as good.

The numbers in last 3 columns represent the percentage of each error category in the model's generation. These are the results of my pilot analysis. For example, the intersection between Hallucination and SGD-DSG represents how many percent of the slot-value pairs generated by the DSG model trained on the SGD dataset are hallucinated (the actual information does not at all exist in the dataset).

Actual Human Evaluation

After deciding on the error categories, Mr. Finch develops an annotation interface for the annotators, who are selected Emory undergraduate students with some exposure to NLP, for them to identify whether the slot-value pair falls into any of the error categories through a procedure similar to the pilot analysis procedure as demonstrated above. The human evaluation result suggests that the end-to-end DSG model trained on the dataset we have developed outperforms the one trained on SGD dataset and the zero-shot GPT pipeline.

Error	Definition	Example Turn with Predicted State Update	SGD-DSG	GPT-Pipe	E2E-DSG

Hallucination	The slot-value represents information that is not mentioned or implied by the corresponding turn at all.	Okay, we'll definitely need to avoid peanuts and lower your sodium intake to manage your blood pressure. Are you currently physically active?
		`activity level: ?`
		`is alcohol served: ?`
			29.0	25.1	34.1

Leakage	The slot-value pair represents information that is only attributable to a previous turn in the dialogue, instead of being derivable from the focal turn.	Alright, Sarah. Can you tell me which payroll period the issues occurred in?
		`payroll period: ?`
		`name: Sarah Johnson`
		`employee ID: 123456`
			2.8	16.0	26.5

Inaccurate	The slot-value pair poorly represents the intended information in the focal turn.	Sure, let's have a sectional rehearsal for that. Bass and cello sections, please come forward. I noticed that in measure 100, some of you were playing a bit too loudly. Please remember to keep the dynamics level until measure 112.
		`beginning note of song: 100`
		`end of song: 112`
			30.8	6.3	12.1

Partial	The value for the particular slot only partially captures the full meaning presented in the focal turn and is thus incomplete.	I believe the target audience for the yoga article would be pregnant women and new mothers, while the sugar article would appeal to a broader group interested in health and wellness. `audience type: Health & Wellness`	12.1	3.3	2.8

Redundant	The slot-value pair represents the same information as another slot-value pair for the focal turn.	My name is John Smith and you can reach me at 555-1234.
		`name: John Smith`
		`contact info: 555-1234`
		`first name: John`
		`last name: Smith`
		`phone number: 555-1234`
			0.9	19.1	12.4

Imprecise	The slot-value representation is unclear and ambiguous in meaning when taken on its own rather than clearly and concisely summarizing some presented information from the focal turn.	When I was growing up, my parents always taught me to appreciate nature and not take it for granted. We used to go on camping trips and nature walks, and those memories stay with me even today.
		`parents teachings: Appreciate nature`
		`camping trips: Yes`
		`nature walks: Yes`
		`camping nature memories: Yes`
			0.9	9.8	15.5

Missing	Some information that is presented in the focal turn and is necessary for capturing its full meaning is not represented by any slot-value pair for the turn.	Well, my current schedule is Monday through Friday, 8:00 am to 5:00 pm. I was thinking about changing it up a bit.
		`day of appointment: Monday`
		`end of the available event: 5:00 pm`
			27.1	0.0	0.0

Good	No errors		30.8	34.7	26.5
Total slot-value pairs predicted			107	398	355

Table 4.2: Elaboration of Error Categories and Model Performances

Chapter 5

Conclusion

Our work has suggest a great significance of conducting data analysis on both the data and on the model output.

For the resume & job description matching task, we have investigated the job description dataset and found that the dataset does not fit the application scenario. This means that even the model suggested promising results after being trained on the previous dataset, the model would not work when actually deployed. Therefore, it is essential to create new datasets when the dataset does not target the application very well.

For the competence-level classification task, our work shows that data quality is very important for training NLP models, and we should conduct error analysis on the data before training the model. Otherwise, even though the model could be capable of handling the task, if the data is noisy, no matter what kind of machine learning we conduct over the data, the model would not work as well as expected. Our work has singlehandedly improved the performance of the BERT model by 5% accuracy through not changing a single bit of the model but just doing data cleaning, suggesting great importance of data analysis and cleaning.

On dialogue state generation, we have conducted data analysis in multiple steps.

When we selected the existing datasets, MultiWOZ [3] and SGD [23], we have looked into these datasets to make sure that they are somehow directly related to our task of dialogue state generation. Then, after we trained the DSG model on these datasets, we looked into the results and found that these does not work very well for DSG due to the limited variety of domains they cover, and therefore we should develop a new dataset for DSG. While developing the dataset, we found through a pilot analysis that if we directly instruct GPT to generate the dialogues from scenario descriptions, GPT would generate dialogues that lacks any useful information, based on which we developed an intermediate step to composite the information before generating the dialogues. After we developed the dataset, we have conducted a round of error analysis, and decided the data was decent. We trained the model over the dataset we have developed, and then we conducted an intensive error analysis on the model output to evaluate the model. In every step, error analysis has been proven essential in guiding the project to the right direction.

To summarize, data analysis is the most important in two of the steps in NLP research: analyzing the data quality and analyzing the model output. Conducting error analysis in these steps are essential in guiding the NLP research project.

Appendix A

Appendix

A.1 CRC1

1. High School Diploma, GED or Program Certificate (CNA, MA, Phlebotomy, Lab Tech) AND 1-year experience in a clinical setting/clinical role.
2. Technical Diploma (LPN, Medical Assistant)
3. Associate Degree or 2 years of college AND 1 year experience in a clinical setting/clinical role.
4. Bachelor's Degree in a scientific or health related field
5. Bachelor's Degree in a non-scientific, non-health related field AND 1 year experience in a clinical setting/clinical role.
6. Bachelor's Degree in a scientific or health related field AND Master's Degree in a non-scientific field.
7. Bachelor's AND Master's Degree in a non-scientific field or health related field AND 1-year experience in a clinical setting/clinical role.
8. Master's Degree in a scientific or health related field.

Refer to Glossary for definitions of Scientific or Health Related Field/ Non-scientific, non-health related field / Clinical setting/clinical role

NOT Recognized as a Clinical setting/clinical role: Business Analyst, Financial Navigator, Massage Therapist

A.2 CRC2

1. High School Diploma, GED or Program Certificate (CNA, MA, Phlebotomy, Lab Tech)
    (a) 1 year experience in a clinical setting/clinical role
    (b) AND 1 year of clinical research experience.
2. Technical Diploma (LPN, Medical Assistant)
    (a) 1 year experience in a clinical setting/clinical role
    (b) AND 1 year of clinical research experience.
3. Associate Degree or 2 years of college
    (a) 1 year experience in a clinical setting/clinical role
    (b) AND 1 year of clinical research experience.
4. Bachelor's Degree in any field AND 1 year of clinical research experience.
5. Master's degree any field AND 1 year clinical research experience
6. Master's of Clinical Research (MSc); no clinical research required
7. MD or PhD in a scientific or health related field (Includes unlicensed US, foreign trained MD's)
8. PhD in a non-scientific or non-health related field AND 1 year clinical research experience

9. Laboratory/Bench Researcher – Laboratory/Bench research required

   (a) Bachelor's in scientific field AND 3 years' research lab experience.

   (b) Master's in scientific field AND 2 years' research lab experience.

   (c) 2 years is comprised of internship or intensive experiences during course work

Clinical research experience (Human subject) is defined as paid employment in clinical research. Residencies and internships will be counted if hours of experience are designated at or above 1000 hours. Otherwise, research experiences during an academic program of study are not considered clinical research experience.

A.3 CRC3

1. High School Diploma, GED or Program Certificate (CNA, MA, Phlebotomy, Lab Tech) AND 3 years of experience in clinical research.
2. Technical Diploma (LPN, Medical Assistant) AND 3 years of experience of clinical research.
3. Associate Degree or 2 years of college AND 3 years of clinical research experience.
4. Bachelor's Degree any field AND 2 years of clinical research experience.
5. Master's Degree any field AND 2 years of clinical research experience
6. Master's of Clinical Research (MSc) AND 1 year of clinical research experience
7. MD or PhD in a scientific or health related field (Includes unlicensed US, foreign trained MD's) AND 1 year of clinical research experience
8. PhD in a non-scientific or non-health related field AND 2 years of clinical research experience
9. Laboratory/Bench Researcher – Laboratory/Bench research required

(a) Bachelor's in scientific field AND 3 years' research lab experience AND 1 year of clinical research (Human Subject)

(b) Master's in scientific field AND 2 years' research lab experience AND 1 year of clinical research (Human Subject)

(c) 2 years is comprised of internship or intensive experiences during course work

Clinical research certification preferred (Required within 1 year of hire date) from a professional organization such as ACRP or SOCRA.

A.4 CRC4

Certification from a professional clinical research organization required on hire

1. High School Diploma, GED or Program Certificate (CNA, MA, Phlebotomy, Lab Tech) AND 5 years of clinical research experience.
2. Technical Diploma (LPN, Medical Assistant) AND 5 years of clinical research experience.
3. Associate Degree or 2 years of college AND 5 years of clinical research experience.
4. Bachelor's Degree AND 4 years of clinical research experience.
5. Master's degree AND 3 years of clinical research experience.
6. Master's of Clinical Research (MSc) AND 2 year of clinical research experience
7. MD or PhD in a scientific or health related field (Includes unlicensed US, foreign trained MD's) 2 years of clinical research experience
8. PhD in a non-scientific or non-health related field AND 3 years of clinical research experience
9. Laboratory/Bench Researcher – Laboratory/Bench research required

(a) Bachelor's in scientific field AND 6 years' research lab experience AND 1 year of clinical research (Human Subject)

(b) Master's in scientific field AND 4 years' research lab experience AND 1 year of clinical research (Human Subject)

These guidelines are to be used for hiring. Promotion is typically based on meeting qualifications for next level positions within institutions with tiered level positions.

A.5 Glossary of Terms

Clinical Setting

- Hospital
- Clinic
- Doctor/Physician Office

Clinical Role*

- Patient Service(s) Coordinator, Patient Care Coordinator, Clinical Coordinator, Unit Secretary,
- Clinical Service Representative,
- Medical Scribe, Medical Secretary,
- Pharmacy Technician,
- Phlebotomist,
- Tumor Registrar,
- Pre/Post Award Administrator,
- Internship in a scientific or health related area (1000 documented hours)

*Semester based academic research experiences noted as less than 6 months are not considered clinical research experience.

Clinical research experience is defined as a minimum of

1. 1 year of paid employment in human subject's clinical research
2. 1 year of Graduate/Undergraduate human subjects research experience which includes residencies and internships

Exclusion: NOT recognized Clinical Role/Setting: Business Analyst, Financial Navigator, and Medical Records Specialist

Bachelor's Degrees in scientific or health related fields*: Biology; Psychology; Epidemiology; Chemistry; Biomedical Science; Neuroscience; Behavioral Science; Sociology; Social Work; Microbiology; Nutrition/Food; Public Health; Health Promotion and Global Health; Complementary and Alternative Health; Community Wellness.

*Licensed Nursing and Licensed Pharmacist are excluded from this list as it would fall under a CRN or Research Pharmacist position.

Bachelor's Degrees in non-scientific or non-health related fields: Business Management; Healthcare Management; Healthcare Administration; Informatics; Gender and Global Health

Clinical Research Roles/Titles

Laboratory Research- Including animal and industry research- nonclinical: laboratory specialist; medical technologist; laboratory assistant; laboratory scientist; some PhD fields are identifiable as laboratory research – chemistry, biotechnology, etc.

Clinical Research Experience: clinical research coordinator; clinical research associate; clinical research assistant; research interviewer; research assistant; clinical research manager/supervisor

Clinical Research Certifications

Society of Clinical Research Associates (SOCRA) OR Association of Clinical Research Professionals (ACRP)

Appendix B

Appendix

B.1 Original Resume Example

John Doe

400 Dowman Drive, GA 30322

(XXX) XXX-XXX

john.doe@email.com

SUMMARY

Highly skilled and accomplished Project Manager with 10+ years of management experience in government and academic settings. Recognized and demonstrated excellence in protocol, instrument, and consent form development, gaining and maintaining IRB clearance, coordinating protocol screening, eligibility, recruitment, enrollment, randomization, and study management. Responsible for coordination of clinical, laboratory, and data activities to ensure compliance with protocols. Accountable for detailed collection and coordination of data entry and quality assurance activities. Excels in creating and maintaining productive collaborative relationships with colleagues both local and long distance. Gap in work history due to being a stay-at-home parent to two children. Currently working at Emory University as Research Administrative Coordinator, Senior.

EDUCATION

- Master of Public Health, Behavioral Science and Health Education. Emory University, The Rollins School of Public Health, Atlanta, GA. (1998)
- Bachelor of Science, Human Development. University of California at Davis, Davis CA. (1992)

PROFESSIONAL EXPERIENCE

Emory University, School of Medicine, Division of Cardiology, Research Administrative Coordinator, Senior. Atlanta GA. (2014-present: Full-time M-F 40/hours/week)

Manages administrative tasks associated with the multi-site research project Patient-Centered Approaches to Research Enrollment Decisions in Acute Cardiovascular Disease (P-CARE). Assists in the development of the study interview guide and other study documents. Submits, gains and maintains IRB and ROC (Grady Hospital) clearance. Ensures the project is administered according to research protocol. Schedules and conducts subject interviews, focus groups, and Patient Advisory Panel meetings. Gathers and manages data. Makes and gives presentations. Serves as the project liaison to multiple study sites, other departments within Emory, outside organizations, government agencies, and product representatives. Assists with grants, expenditure monitoring, and budgeting. Contributes to data analysis. Performs additional related responsibilities as required.

Additionally, acts as Administrative Assistant for Emory Clinical Cardiovascular Research Institute (ECCRI). Processes purchase orders, requisitions, check requests, invoices, and expense reports via Compass or Emory Express for junior research faculty and cardiology research fellows. Provides calendar support. Purchases supplies for various research studies, as needed. Works closely with various departments and vendors to solve problems and facilitate the timely delivery of goods and services. Works with Cardiology fellows applicants during the interview process and assists with onboarding.

Life Science Partner, Research Associate, Atlanta GA. (2014: Part-time 20/hours/week)

Life Science Partner is a national recruiting firm dedicated to providing clients with top candidates from within the Medical, Health, Pharma, and Biotech industries. Key responsibilities include candidate identification – focused and frequent research for a richer, fuller pipeline of candidates and management of email communications. Proofing, quality control, filing of internal and external documents. Manages multiple databases. Works closely with and assists the Vice President.

Centers for Disease Control and Prevention (CDC)/Emory University, Project Manager, Women With Bleeding Disorders Study, Atlanta, GA. (1999-2003: Full-time 40/hours/week)

Oversaw the management of a 7-site nation-wide prevention intervention study to assess different treatments for menorrhagia and their effect on blood loss and quality of life. Researched and selected quality of life data collection instruments. Acquired loaned Chrono-log Aggregometers from Chrono-log Corporation. Negotiated with Aventis Behring and Pharmacia & Upjohn pharmaceutical companies to acquire donated study drug. Gained acquisition of Investigational New Drug (IND) status from the Food and Drug Administration (FDA) for one study drug. Assisted in database development and maintenance. Served as the point-person for all sites staff and CDC.

Emory University Research Associate Senior, Project FAST (Female Atlanta STudy), Atlanta GA. (1998-1999: Full time 40/hours/week)

Project FAST is a study of multigenerational drug use among women. Coordinated recruitment, screening, and interviewing of study participants. Organized and cleaned data. Macro coded qualitative interviews. Analyzed data for presentation and publication.

Emory University, Graduate Research Assistant, Project FAST, Atlanta, GA. (1996-1998: Part-time 20/hours/week)

Developed qualitative interview guide. Established community contacts and social networks. Recruited study participants and conducted quantitative and qualitative interviews. Generated data analysis program using QSR NUD*IST.

SELECTED CONSULTING/COLLABORATIONS

- National Hemophilia Foundation. Report: Availability of Adolescent/Teen Programming at Hemophilia Treatment Centers and Chapters. (2001)
- National Hemophilia Foundation. Report: Collaborations Between Centers of Excellence in Women's Health and Hemophilia Treatment Centers. (2001)
- National Hemophilia Foundation. Grant Reviewer, The National Prevention Program. Preventing Complications: Strategies for Youth With Hemophilia.
- Emory University. Recruitment Coordinator, Lesbian Sex Project. (2001)
- National Immunization Coalition. (1998-1999) Consultant and Focus Group Facilitator. (1998)

COMMUNITY SERVICE

- Vice President Board of Directors. The Inn Between: A Transitional Shelter for Women and Their Children, Atlanta, GA. (1998-2000)
- Co-Facilitator Women's Support Group. Santa Fe Villas, Atlanta, GA. AIDS Outreach Educator. Atlanta Harm Reduction Center. (1998-1999)
- Atlanta, GA. Names Project AIDS Memorial Quilt Volunteer. (1996-1998) Sacramento Names Project Chapter, Sacramento, CA. (1993-1996)

- Volunteer Peer Counselor. The House, Peer Counseling Center, University of California at Davis, Davis, CA. (1989-1992)

PUBLICATIONS

Jane Doe and John Doe. (1998). Atlanta: Metropolitan Atlanta Drug Abuse Trend. In Epidemiologic Trends in Drug Abuse (NIH Publication, 1998) Washington, DC: U.S. Department of Health and Human Services.

TRAINING

- Tuberculosis Update and Tuberculin Skin Test Certification Workshop. Georgia Tuberculosis Control Program, Atlanta GA. (1997)
- HIV Antibody Counseling. California State Office of AIDS, Sacramento, CA. (1993)
- Community Health Outreach Worker Training. Yolo County Department of Health, Woodland, CA. (1993)

COMPUTER SKILLS

Proficient with Emory Express, Compass, Outlook, WordPerfect, Microsoft Word, PowerPoint, QSR NUD*IST, Excel, and SPSS. Familiar with EpiInfo and SAS.

B.2 Concatenated Input

> Master of Public Health, Behavioral Science and Health Education. Emory University, The Rollins School of Public Health, Atlanta, GA. (1998) Bachelor of Science, Human Development. University of California at Davis, Davis CA. (1992) ¡sep¿ ¡sep¿ Emory University, School of Medicine, Division of Cardiology, Research Administrative Coordinator, Senior. Atlanta GA. (2014- present : Full-time M-F 40/hours / week) Manages administrative tasks associated with the multi-site research project Patient-Centered Approaches to Research Enrollment Decisions in Acute Cardiovascular Disease (P-CARE) . Assists in the development of the study interview guide and other study documents.

Submits, gains and maintains IRB and ROC (Grady Hospital) clearance. Ensures project is administered according to research protocol. Schedules and conducts subject interviews, focus groups, and Patient Advisory Panel meetings. Gathers and manages data. Makes and gives presentations. Serves as project liaison to multiple study sites, other departments within Emory, outside organizations, government agencies and product representatives. Assists with grants, expenditure monitoring and budgeting. Contributes to data analysis. Performs additional related responsibilities as required. Additionally, acts as Administrative Assistant for Emory Clinical Cardiovascular Research Institute (ECCRI) . Processes purchase orders, requisitions, check requests, invoices, and expense reports via Compass or Emory Express for junior research faculty and cardiology research fellows. Provides calendar support. Purchases supplies for various research studies, as needed. Works closely with various departments and vendors to solve problems and facilitate timely delivery of good and services. Works with Cardiology fellows applicants during the interview process and assists with onboarding. Life Science Partner, Research Associate, Atlanta GA. (2014 : Part-time 20/hours / week) Life Science Partner is a national recruiting firm dedicated to providing clients with top candidates from within the Medical, Health, Pharma and Biotech industries. Key responsibilities include candidate identification - focused and frequent research for richer, fuller pipeline of candidates and management of email communications. Proofing, quality control, filing of internal and external documents. Manages multiple databases. Works closely with and assists the Vice President. Centers for Disease Control and Prevention (CDC) /Emory University, Project Manager, Women With Bleeding Disorders Study, Atlanta, GA. (1999-2003 : Full-time 40/hours / week) Oversaw the management of a 7-site nation-wide prevention intervention study to assess different treatments for menorrhagia and their effect on blood loss and quality of life. Researched and selected quality of life data collection instruments. Acquired loaned Chrono-log Aggregometers from Chrono-log Corporation. Negotiated with Aventis Behring and Pharmacia & Upjohn pharmaceutical companies to acquire donated study drug. Gained acquisition of Investigational New Drug (IND) status from the Food and Drug Administration (FDA) for one study drug. Assisted in data base development and maintenance. Served as point-person for all sites staff and CDC. Emory University Research Associate Senior, Project FAST (Female Atlanta STudy) , Atlanta GA. (1998-1999 : Full time 40/hours / week) Project FAST is a study of multigenerational drug use among women. Coordinated recruitment, screening and interviewing of study participants. Organized and cleaned data. Macro coded qualitative interviews. Analyzed data for presentation and publication. Emory University, Graduate Research Assistant, Project FAST, Atlanta, GA. (1996-1998 : Part-time 20/hours / week) Developed qualitative interview guide. Established community contacts and social networks. Recruited study participants and conducted quantitative and qualitative interviews. Generated data analysis program using QSR NUD IST. SELECTED CONSULTING / COLLABORATIONS National Hemophilia Foundation. Report : Availability of Adolescent / Teen Programming at Hemophilia Treatment Centers and Chapters. (2001) National Hemophilia Foundation. Report : Collaborations

Between Centers of Excellence in Women's Health and Hemophilia Treatment Centers. (2001) National Hemophilia Foundation. Grant Reviewer, The National Prevention Program. Preventing Complications : Strategies for Youth With Hemophilia. Emory University. Recruitment Coordinator, Lesbian Sex Project. (2001) National Immunization Coalition. (1998-1999) Consultant and Focus Group Facilitator. (1998) COMMUNITY SERVICE Vice President Board of Directors. The Inn Between : A Transitional Shelter for Women and Their Children, Atlanta, GA. (1998-2000) Co-Facilitator Women's Support Group. Santa Fe Villas, Atlanta, GA. AIDS Outreach Educator. Atlanta Harm Reduction Center. (1998-1999) Atlanta, GA. Names Project AIDS Memorial Quilt Volunteer. (1996-1998) Sacramento Names Project Chapter, Sacramento, CA. (1993-1996) Volunteer Peer Counselor. The House, Peer Counseling Center, University of California at Davis, Davis, CA. (1989-1992) ¡sep¿ Research Administrative Coordinator

Appendix C

Appendix

Task Overview

1. Checking Filled Slots

   - Slot Names:

     - Slot names should describe a type of information, not a specific value.
     - Ensure that slot names align with the information present in the turn.
     - Modify slot names if they are too specific or not reflective of the information.
     - If a slot name cannot be fixed, remove the slot.

   - Slot Values:

     - Slot values should be mentioned or strongly implied in the turn.
     - Verify that slot values match the corresponding slot names.
     - Modify slot values if they do not align with the information.
     - If a slot value cannot be fixed, remove the slot.

   - Slot Descriptions:

     - Slot descriptions should reasonably represent the slot name and value.
     - Ensure that slot descriptions do not explicitly reveal the slot value.
     - Modify slot descriptions if they give away the slot value.

2. Adding Missing Information

   - Check if there is any information that is missing but explicitly or coreferentially mentioned in the turn.
   - Only add a slot if you are confident that the slot is correct and directly related to the turn's content.
   - Add missing slots by creating a new row below the existing filled slots (above the unfilled slots).

3. Reviewing Unfilled Slots

   - Unfilled slots should describe information NOT present in the turn.
   - Unfilled slots should be loosely related to the topic of conversation.
   - No need to assess the reasonability of unfilled slot names or descriptions.

Guidelines for Editing The CSV

- Slot Modification

  – To modify a slot, separate the existing value with two pipes (||) and add the new value after the pipe(s).

  Example: "Old Value || New Value"

- Slot Removal

  – If a slot name, value, or description cannot be fixed, delete the row containing that slot.

- Adding Missing Slots

  – Ensure that missing slots have a name, value, and description.

  – Add missing slots in a new row below the existing filled slots, above unfilled slots.

Example Workflow

1. Review the turn's content.

2. Check if filled slots are correct (name, value, description).
3. Modify or remove slots as needed.
4. Check for missing information and add slots if appropriate.
5. Review unfilled slots for relevance.

Additional Notes

- Maintain consistency in slot naming conventions.
- Focus on accuracy, clarity, and concision in slot descriptions.
- Avoid introducing new information or making assumptions not supported by the dialogue.
- If a turn does not contain any relevant information for slot filling, no action is required for that turn.

Bibliography

[1] Irfan Ali, Nimra Mughal, Zahid Hussain Khand, Javed Ahmed, and Ghulam Mujtaba. Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal Of Engineering & Technology*, 41(1):65–79, 2022. URL https://search.informit.org/doi/10.3316/informit.263278216314684.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0387310732.

[3] Pawe\l Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

[4] Derek Chen, Kun Qian, and Zhou Yu. Stabilized In-Context Learning with Pre-trained Language Models for Few Shot Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1551–1564, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.115.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[6] Gaurav Dutta Kumar. Resume dataset. Kaggle, 2024. `https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset`.

[7] Han He and Jinho D. Choi. Unleashing the true potential of sequence-to-sequence models for sequence tagging and structure parsing. *Transactions of the Association for Computational Linguistics*, 11:582–599, 2023. doi: 10.1162/tacl_a_00557. URL `https://aclanthology.org/2023.tacl-1.34`.

[8] Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gašić. ChatGPT for Zero-shot Dialogue State Tracking: A Solution or an Opportunity?, June 2023. URL `http://arxiv.org/abs/2306.01386`. arXiv:2306.01386 [cs].

[9] Kemmogne Fofana Fabrice Jiechieu and Ngnotue Tsopze. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing & Applications*, 33:5069–5087, 2021. doi: 10.1007/s00521-020-05302-x.

[10] Brendan King and Jeffrey Flanigan. Diverse Retrieval-Augmented In-Context Learning for Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5570–5585, Toronto, Canada, 2023. Asso-

ciation for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.344. URL `https://aclanthology.org/2023.findings-acl.344`.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[12] Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D. Choi. Competence-level prediction and resume & job description matching using context-aware transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8456–8466, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.679. URL `https://aclanthology.org/2020.emnlp-main.679`.

[13] Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D. Choi. Competence-level prediction and resume & job description matching using context-aware transformer models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8456–8466, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.679. URL `https://aclanthology.org/2020.emnlp-main.679`.

[14] Renxuan Albert Li, Ihab Hajjar, Felicia Goldstein, and Jinho D. Choi. Analysis of hierarchical multi-content text classification model on B-SHARP dataset for early detection of Alzheimer's disease. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 358–365, Suzhou, China,

December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.38.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages –, 2019.

[16] Spoorthi M, Indu Priya B, Meghana Kuppala, Vaishnavi Sunilkumar Karpe, and Divya Dharavath. Automated resume classification system using ensemble learning. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1782–1785, 2023. doi: 10.1109/ICACCS57279.2023.10112917.

[17] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. *Advances in Neural Information Processing Systems*, 35:462–477, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0346c148ba1c21c6b4780a961ea141dc-Abstract-Conference.html.

[18] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, 1st edition, 2012. ISBN 978-0262018029.

[19] Shabna Nasser, C Sreejith, and M Irshad. Convolutional neural network with word embedding based approach for resume classification. In *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, pages 1–6, 2018. doi: 10.1109/ICETIETR.2018.8529097.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improv-

ing language understanding by generative pretraining. In *Advances in Neural Information Processing Systems*, pages 8165–8175, 2018.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Shruti Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[22] Ravindra Singh Rana. Job description dataset. Kaggle, 2024. `https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset`.

[23] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i05.6394. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6394`. Number: 05.

[24] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL `https://aclanthology.org/2021.eacl-main.24`.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[26] Qingyue Wang, Liang Ding, Yanan Cao, Yibing Zhan, Zheng Lin, Shi Wang,

Dacheng Tao, and Li Guo. Divide, Conquer, and Combine: Mixture of Semantic-Independent Experts for Zero-Shot Dialogue State Tracking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2048–2061, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.114. URL `https://aclanthology.org/2023.acl-long.114`.

[27] Seungpil Won, Heeyoung Kwak, Joongbo Shin, Janghoon Han, and Kyomin Jung. BREAK: Breaking the Dialogue State Tracking Barrier with Beam Search and Re-ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2832–2846, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.159. URL `https://aclanthology.org/2023.acl-long.159`.

[28] Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. MultiWOZ 2.4: A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK, September 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.sigdial-1.34`.