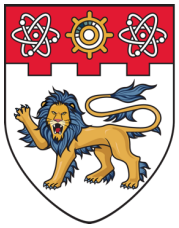
Advancing Low Resource Information Extraction and Dialogue System using Data Efficient Methods



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Ding Bosheng

School of Computer Science and Engineering (SCSE)

Alibaba Group

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

25/01/2024

Date

Ding Bosheng

Ding Bosheng

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

25/01/2024

Date

Assistant Prof. Luu Anh Tuan

Authorship Attribution Statement

This thesis contains material from 3 paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as a first author or co-first author.

Chapter 3 is published as DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. Authors: Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020.

The contributions of the co-authors are as follows:

- I initiated the concept of a semi-supervised, generation-based data augmentation methodology. This involved partaking in the code development, orchestrating and executing experimental designs, and drafting the initial manuscript. Additionally, I collaborated in developing the evaluation code, co-conducting experiments, and contributing to manuscript revisions.
- Linlin engaged in periodic deliberations with Dr. Lidong Bing and myself, playing a pivotal role in refining the manuscript.
- Dr. Lidong Bing participated in regular dialogues with both Linlin and myself, offering vital insights and contributing to manuscript enhancements.
- Dr. Canasai Kruengkrai was instrumental in setting the foundational research trajectory, taking part in coding efforts, and actively involved in manuscript refinement.
- Dr. Thien Hai Nguyen, Prof. Shafiq Joty, Prof. Luo Si, and Prof. Chunyan Miao were key in providing expert advice and played a substantial role in the iterative process of manuscript revision.

Chapter 4 is published as GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. Authors: Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022.

The contributions of the co-authors are as follows:

- The concept of creating a multilingual, task-oriented dialogue system was initially proposed by me, alongside the notion of achieving globalization through the process of localization.
- Prof. Junjie Hu offered mentorship in the development of experimental designs and the composition of scholarly papers.

- Dr. Lidong Bing actively participated in ongoing scholarly discussions.
- Mahani Aljunied contributed to the project by engaging in data annotation and overseeing the quality control process.
- Prof. Shafiq Joty played a pivotal role in both the validation of the research concept and the review process of the academic paper.
- Prof.s Luo Si and Chunyan Miao were instrumental in offering critical feedback and suggestions, which were crucial in the enhancement and subsequent revision of the manuscript.

Chapter 5 is published as Is GPT-3 a Good Data Annotator?. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. Authors: Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023.

The contributions of the co-authors are as follows:

- I initiated the concept of employing Large Language Models (LLMs) for the purpose of data annotation. This encompassed the development of relevant coding, the execution of experimental procedures, and the conduct of interviews with linguists.
- Regular discussions were held with Chengwei Qin, who also undertook the responsibility of conducting experiments pertaining to the SST-2 dataset.
- Linlin Liu contributed significantly to the validation of the experimental setup.
- Yew Chen Chia played a pivotal role in the development of the code and offered valuable insights for enhancing the presentations.
- The significant contributions of Dr. Lidong Bing, Prof. Shafiq Joty, and Prof. Boyang Li were instrumental in offering specialized guidance. Their expertise played a crucial role in the iterative process of refining the manuscript.

25/01/2024

Date

Ding Bosheng

Acknowledgements

My doctoral journey, an extensive and intellectually demanding endeavor, was made possible through the collaborative efforts and support of numerous individuals. Foremost, I wish to express my profound appreciation to my mentor, Prof. Shafiq Joty, whose guidance and mentorship have been invaluable. Prof. Joty not only provided me with an exceptional opportunity to explore the depths of natural language processing, but his expertise and fervent enthusiasm for machine learning have profoundly shaped my professional methodology and approach. The privilege of being his student has been a pivotal and enriching experience in my academic journey.

In addition, my heartfelt gratitude extends to Prof. Luu Anh Tuan. His unwavering support and consistent encouragement were instrumental in navigating the complexities of my Ph.D. studies. The success of my doctoral journey is, in no small part, due to his invaluable assistance.

Moreover, I am immensely thankful to the members of my dissertation committee. Their insightful feedback and constructive criticism played a crucial role in refining my research and scholarly pursuits.

Prof. Junjie Hu also deserves recognition for his extensive research guidance and for imparting wisdom that transcends the academic sphere. Similarly, I am indebted to Prof. Chunyan Miao for her mentorship and support throughout this journey.

I would also like to acknowledge my esteemed colleagues from the NTU-NLP group. Their collaborative spirit and intellectual camaraderie have significantly contributed to my research endeavors. Special thanks are extended to Dr. Liu Linlin, Dr. Lin Xiang, Dr. Jwalapuram Prathyusha, Dr. Maruf Saiful Bari, Li Xingxuan, Qin Chengwei, Jiao Fangkai, Zhao Ruochen, Han Simeng, Chen Hailin, and Wang Weishi for their support and contributions.

Additionally, I am grateful for the camaraderie and support of my colleagues and friends at Alibaba Group. In particular, I extend my sincere appreciation to Ms. Mahani Aljunied for her exceptional collegiality, and to Dr. Haiyun Peng, Dr. Ruidan He, Mr. Zhongwei Li, Dr. Zhen Hai, Dr. Liying Cheng, Guizhen Chen, Chia Yew Ken, Qingyu Tan, Zhaodonghui Li, and Dr. Lidong Bing for their collaboration and shared moments of professional growth. The time spent working together in our lab, meeting challenging deadlines, and the enjoyable moments we shared, have been integral to my personal and professional development.

Undertaking the journey to earn a Doctor of Philosophy (PhD) degree is a challenging and arduous endeavor. The process demands not only intellectual rigor but also a significant commitment of time and energy. In this context, the support of one's personal network is invaluable. I am profoundly grateful for the unwavering support and encouragement provided by my girlfriend, Dr. Jiayi Huang. Her invaluable insights, coupled with her steadfast presence, have been a cornerstone of my success in navigating the complexities and demands of PhD studies. Dr. Huang's guidance and understanding played a pivotal role in my academic journey, offering both professional perspective and personal comfort during the most challenging phases of my doctoral work.

Finally, I wish to express my profound gratitude for the unwavering support provided by my family throughout the course of my academic journey. During the challenging period of my third year in the pursuit of my Doctor of Philosophy degree, I experienced the profound loss of my father. His passing was a significant event in my life, yet the enduring memory of his love and encouragement continues to inspire and guide me. His influence has been a cornerstone in my development both personally and academically, providing me with the strength and resilience to persevere in my studies. This thesis not only stands as a testament to my academic endeavors but also as a tribute to his lasting impact on my life.

In summary, this journey, though individual in nature, was enriched and made successful through the collective efforts and support of these remarkable individuals. Their contributions to my academic and personal growth have been immeasurable, and for this, I remain eternally grateful.

"There is only one true heroism in the world: to see the world as it is, and to love it."

— Roman Rolland

To my parents,
Kunjing Ding and Lifen Huang

Abstract

This thesis, titled presents an extensive study aimed at improving the efficacy of language models in situations characterized by limited data resources, a prevalent challenge in the field of natural language processing (NLP). The research emphasizes the development and refinement of data-efficient methods, which are essential for enhancing the robustness and functionality of language models in environments with scarce data resources.

At the heart of this thesis is the investigation of novel data augmentation approaches designed to enrich the training dataset. These include the creation of synthetic data through advanced algorithms, which generate realistic and varied linguistic examples to augment the training corpus without necessitating manual data annotation. Additionally, the study introduces techniques for semantic data transformation that modify existing data in semantically meaningful ways, thereby exposing models to a diverse range of linguistic structures and contexts.

The research also addresses the utilization of these data augmentation methods to improve language models' resilience to overfitting, a frequent issue in low-resource settings. By diversifying and enriching the training dataset, the models achieve enhanced generalization capabilities, resulting in improved performance on new, unseen data.

Further, the thesis explores the integration of these data augmentation techniques with current NLP models, highlighting the synergistic advantages of combining innovative data enrichment methods with cutting-edge language models. This integration not only increases model robustness but also broadens the models' applicability to a more diverse array of languages and dialects, especially those with sparse data.

Moreover, in the era of Large Language Models (LLMs), this thesis explores algorithms that leverage LLMs' intrinsic abilities to comprehend and generate contextually appropriate augmentations, thus enriching training data while maintaining its quality.

The empirical results presented in this thesis demonstrate the effectiveness of the proposed data augmentation techniques. These results reveal substantial enhancements in model accuracy, resilience, and generalization across various NLP tasks, including sentiment analysis, named entity recognition, part of speech tagging, relation extraction, and task-oriented dialogue systems.

In summary, this thesis makes a significant contribution to NLP by introducing innovative data-efficient methods that bolster the resilience of language models in low-resource scenarios. The research findings and methodologies pave the way for future studies in enhancing language model robustness, thereby expanding the reach of NLP technologies to a broader spectrum of languages and applications. The thesis concludes by identifying and discussing several promising avenues for future research in this domain.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
ASTE	Aspect Sentiment Triplet Extraction
BPE	Byte Pair Encoding
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CV	Computer Vision
DA	Data Augmentation
DST	Dialogue State Tracking
E2E-TBSA	End-to-End Target Based Sentiment Analysis
FFN	Feed Forward Network
FSL	Few-shot Learning
GAN	Generative Adversarial Network
GPU	Graphic Processing Unit
KB	Knowledge Base
KG	Knowledge Graph
KD	Knowledge Distillation
KL	Kullback-Leibler
K-NN	K-Nearest Neighbors
LM	Language Model
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
MLM	Masked Language Modeling
MT	Machine Translation

MTL	Multi-Task Learning
NER	Named Entity Recognition
NLP	Natual Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
NN	Neural Network
PCA	Principal Component Analysis
POS	Part-of-Speech
PLM	Pretrained Language Model
QA	Question Answering
RE	Relation Extraction
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SOTA	State Of The Art
TTS	Text-to-Speech
TOD	Task-Oriented Diaglogue
XAI	Explainable Artificial Intelligence
XLM	Cross-lingual Language Model
ZSL	Zero-shot Learning

e.g.	exemplum gratia (en: for example)
et al.	et alia (en: and others)
i.e.	id est (en: that is)
i.i.d.	independent and identically distributed

Chapter 1

Introduction

The realm of artificial intelligence has witnessed a paradigm shift with the advent and proliferation of deep learning [1]. At the core of this revolution are neural network models [2, 3], intricate architectures composed of multiple layers of nonlinear functions. These models, drawing inspiration from the biological neurons in animal brains [4], have demonstrated unparalleled proficiency in modeling complex functions and phenomena [5]. Their versatility and expressiveness have been pivotal in their widespread application across various domains, including natural language processing (NLP) [6], computer vision (CV) [7], speech recognition [8], and even drug discovery [9]. Such expansive applications underscore the transformative impact of these neural network models in the landscape of computational technology and research.

The past decade has seen significant strides in computer hardware, particularly in the application of graphic processing units (GPUs), enabling more efficient training of large-scale neural models. These advancements have led neural model-based methods to achieve state-of-the-art performance in a multitude of tasks. Unlike traditional statistical methods, which rely heavily on manually crafted features and rules [10, 11], neural models function as automatic feature extractors, adept at distilling relevant information from complex datasets [12]. This shift towards automation has redefined the approaches to information processing, making neural models an invaluable tool in the arsenal of modern computing.

However, the high performance of neural models comes with its own set of challenges [13]. Their tendency to be over-parameterized makes them susceptible to

overfitting, particularly when exposed to noise or irrelevant features [14]. This over-parameterization necessitates a substantial volume of labeled training data to optimize performance. The collection and annotation of such data is often a resource-intensive process, both in terms of time and cost. This dependency on large datasets poses significant hurdles in low-resource scenarios, where data availability is limited.

Recognizing this challenge, the research community has increasingly focused on enhancing the robustness of neural models in low-resource settings [15]. Several strategies have been proposed to mitigate the risk of overfitting, including the implementation of various regularization techniques, the adoption of semi-supervised learning approaches to utilize unlabeled data more effectively, and the application of transfer learning to leverage knowledge from other datasets or tasks [16].

This thesis aims to contribute to this burgeoning field of study. We explore innovative methods to enhance the robustness and efficiency of neural models, particularly in the context of natural language processing under low-resource conditions. Our objective is to develop techniques that enable these models to maintain high performance with minimal reliance on extensive labeled datasets, thereby paving the way for more accessible and sustainable applications in various domains. The forthcoming chapters will delve into the specifics of these methods, their implementation, and the impact they promise in advancing the field of AI in low-resource environments.

In the subsequent sections of this introduction, we will initially explore the driving forces and goals behind our study (Section 1.1), followed by a brief overview of the key achievements from our finished research projects (Section 1.2). To conclude, the structure of this thesis will be delineated in Section 1.3.

1.1 Motivations and Objectives

As previously discussed, there is a wide array of strategies put forward in past research aimed at decreasing dependence on extensive datasets and enhancing the resilience of neural models in environments with limited resources [17]. These strategies predominantly encompass techniques to minimize overfitting, as well as semi-supervised and transfer learning approaches. Methods to mitigate overfitting

can be further divided into categories such as data augmentation, label smoothing, regulating model parameters, controlling hidden representations, hybrid methods, among others.

Data augmentation has emerged as a notably effective method, particularly in the realm of image processing[13]. Nonetheless, its application in Natural Language Processing (NLP) poses greater challenges due to its vulnerability to noise and the potential for even minor word replacements to alter an entire sentence's meaning [16]. Therefore, the development of more efficient data augmentation techniques for NLP, especially for detailed tasks like sequence tagging, is urgently needed. Moreover, as fine-tuning pre-trained neural models becomes increasingly standard for various natural language and image processing tasks, new challenges arise in enhancing the performance of these models in low-resource settings [18]. In this thesis, we aim to tackle the following research questions (RQ) using the methods we propose:

RQ1: What are effective ways to augment training data for sequence tagging tasks while maintaining consistency in tokens and tags?

In the field of Natural Language Processing (NLP), particularly in specialized tasks such as sequence tagging, the process of applying data augmentation proves to be exceptionally challenging. This challenge largely stems from the high sensitivity of these tasks to any form of noise or inconsistency. Traditional approaches to data augmentation, including but not limited to back-translation [19], as well as random processes of deletion, insertion, or substitution of words within sentences [20], have shown better compatibility with tasks at the sentence level rather than at the sequence tagging level. The reason for this is that these methods, when applied to sequence tagging tasks, can lead to unpredictable and often undesirable alterations in both the sequence and nature of the tags assigned to the data.

In the specific context of this research, detailed in Chapter 3, we delve into analyzing the limitations and inadequacies inherent in these prevalent methods. Moreover, we put forth new and inventive generation-based techniques that are tailored for sequence tagging. These novel methods primarily leverage advanced language models. These models are adept at generating synthetic, yet realistically labeled, sequences. This approach to data augmentation not only enriches the training data with a greater diversity of examples but also addresses the critical issue of

maintaining the integral consistency between tokens and their respective tags. By integrating these innovative methods, we aim to substantially improve the quality and effectiveness of the training data used in sequence tagging tasks, thereby enhancing the overall performance of models trained on such data.

RQ2: What are the advanced methods for creating a multilingual task-oriented dialogue system that effectively integrates local entities and context-specific nuances within the framework of global communication?

The advancement of task-oriented dialogue systems has predominantly concentrated on languages that have a wealth of resources [21]. This focus has led to significant progress in these languages. However, an aspect that has not received adequate attention is the localization of these systems in a multilingual context [22]. The lack of exploration in this area has constrained the practical deployment and effectiveness of multilingual task-oriented dialogue systems in diverse linguistic and cultural settings.

In the realm of global communication, the incorporation of local entities and the understanding of context-specific nuances are crucial for the successful operation of these systems. The need for such systems to be versatile and adaptable in handling various languages and dialects, each with their own unique characteristics and cultural contexts, is more pressing than ever.

In Chapter 4, we delve into this under-researched area by examining the influence of local entities and cultural context on the performance and reliability of multilingual task-oriented dialogue systems. We explore how these systems can be optimized to recognize and respond to local specifics, such as regional language variants, culturally relevant terms, and locally significant concepts, which are often overlooked in the current models.

Moreover, we propose an innovative yet straightforward methodology for developing these systems. Our approach focuses on enhancing their ability to function effectively in a multilingual and multicultural landscape, particularly in the context of globalization. This involves creating systems that are not only linguistically diverse but also culturally sensitive, ensuring that they can operate efficiently across different geographical regions and cultural backgrounds.

By addressing these aspects, we aim to bridge the gap in the current understanding and implementation of multilingual task-oriented dialogue systems, paving the way for more inclusive and effective global communication. This research is expected to contribute significantly to the field, offering new insights and practical strategies for developing dialogue systems that are truly global in their reach and application.

RQ3: In what ways can advanced large language models (LLMs) fundamentally transform the methodologies and enhance the effectiveness of data annotation processes across a wide range of sectors?

In the current landscape, the process of data annotation is heavily reliant on human input, making it both time-consuming and expensive. This reliance presents significant challenges, not only for individuals but also for research institutions and small-to-medium enterprises (SMEs) who may find the costs prohibitive [23]. Chapter 5 delves into an in-depth examination of how large language models can be utilized as a tool for data annotation. This exploration includes a detailed comparison between the outcomes and associated costs of annotations performed by LLMs versus those conducted by human annotators.

Moreover, this research broadens its scope to assess the versatility of LLMs in various specialized tasks where data annotation plays a pivotal role. These tasks include, but are not limited to, sentiment analysis, named entity recognition (NER), relationship extraction, and the intricate process of aspect sentiment triplet extraction. By examining these areas, the study aims to provide comprehensive insights into the potential of LLMs to not only streamline the data annotation process but also to enhance its accuracy and efficiency across diverse domains. This extensive exploration could pave the way for a new era in data annotation, marked by increased automation and reduced dependency on human labor.

1.2 Main Contributions

In this section, we outline the key contributions our research makes to the advancement of robust methodologies in natural language processing, particularly in resource-constrained environments. Our study introduces a comprehensive array of solutions aimed at addressing the identified challenges, covering several aspects.

Key among these are: the adoption of data augmentation strategies, the enhancement of pre-existing language models for improved cross-lingual transfer learning, the employment of advanced large language models in data distillation processes to better utilize augmented data, and the fusion of modern neural network models with traditional techniques to reduce reliance on extensive training datasets. These innovative approaches not only tackle current issues but also establish a foundation for subsequent research in this field.

1.2.1 Generation-based data augmentation methods for fine-grained Information Extraction Tasks

In our research, we place a strong emphasis on the importance of diversifying original training data as a vital tactic to reduce the risk of overfitting in machine learning models. This is particularly crucial when addressing token-level Natural Language Processing (NLP) tasks, such as sequence tagging. These tasks are uniquely challenging due to their heightened sensitivity to any noise that might be introduced during the data augmentation process. To address this, we have diligently conducted comprehensive studies that delve into the nuances of data augmentation specifically tailored for sequence tagging.

Chapter 3 unveils a cutting-edge data augmentation technique, designed specifically for an array of sequence tagging tasks. This novel approach starts with an intricate process of transforming both the sentence tokens and their corresponding labels into a linearized format. This is achieved by placing each label directly before its associated token. In doing so, the labels assume the role of modifiers in the sequence. Following this reconfiguration, we engage language models trained on these linearized label sequences. This training is vital as it enables the models to grasp the underlying distribution of both tokens and labels within the data.

Once this understanding is established, our technique advances to the next phase, which involves the generation of additional synthetic sequences. This is accomplished by utilizing a specially designated token at the start of each sequence. This innovative methodology serves a dual purpose. Firstly, it significantly broadens the variety of the training data, an essential step for enhancing the robustness of the models. Secondly, it incorporates elements of a semi-supervised learning approach. This is done by making use of unlabeled data as well as data that has been labeled

through dictionaries. The inclusion of these diverse data types contributes to a more comprehensive training process.

Our approach has demonstrated consistent and noteworthy improvements across several key areas. These include Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and End-to-End Target-Based Sentiment Analysis (E2E-TBSA). These improvements are a testament to the efficacy of our data augmentation technique in enhancing the performance of models on these specific tasks.

To foster collaborative progress and transparency in the field, we have made the source code of our innovative technique publicly accessible. Interested parties, researchers, and practitioners in the field can access and utilize this code by visiting our GitHub repository at https://github.com/ntunlp/daga.git. This open-source initiative underscores our commitment to contributing to the collective advancement of NLP research and development.

1.2.2 Developing a Task-Specific Multilingual Dialogue System Incorporating Local Entities and Contexts in a Global Setting

Chapter 4 explores the intricate development of a specialized multilingual dialogue system. This system is uniquely constructed to incorporate local entities and contexts while operating within a broader, global framework. Its primary objective is to adeptly handle and react to a wide array of user queries. This capability is especially beneficial in settings where having localized knowledge and the ability to communicate in multiple languages is essential.

The essence of our approach is centered around devising a resilient model capable of comprehending and generating replies in several languages. A critical aspect of this model is its heightened sensitivity to local subtleties, specific entities, and contextual details. The chapter places a particular emphasis on how local entities and contexts are incorporated into the creation of this multilingual, task-oriented dialogue system, aimed at serving global purposes.

To bring this vision to fruition, we adopt a combined approach, leveraging cutting-edge natural language processing methodologies alongside a sophisticated deep

learning framework. Our system undergoes rigorous training with an extensive dataset that includes a variety of languages and regional dialects. This dataset is further enriched with an ample amount of localized content, ensuring the system's proficiency in language translation and in the nuanced understanding and integration of local cultural and contextual factors.

The preliminary outcomes of our research show a notable improvement in the system's proficiency to offer precise and contextually appropriate responses in a multilingual environment. This indicates the system's considerable potential as a valuable asset in global applications where local relevance is paramount. Through this chapter, we aim to provide a comprehensive understanding of the processes and strategies involved in building such a system, highlighting its effectiveness and the technological advancements that make it possible.

In our continued efforts to drive collective innovation and maintain openness in our research endeavors, we are pleased to announce that the source code for our cutting-edge methodology is now openly available to the public. Enthusiasts, scholars, and professionals in the relevant field are encouraged to explore and employ this code, which can be found on our GitHub repository at https://github.com/Bosheng2020-/globalwoz. This move towards open-sourcing is a testament to our dedication to fostering the ongoing growth and progress in the realm of NLP research and application.

1.2.3 Utilizing Large Language Models for Enhanced Data Labeling in Natural Language Processing Tasks

Chapter 5 of our study offers a comprehensive exploration into the innovative use of Large Language Models (LLMs) for the purpose of automatic data annotation in the field of Natural Language Processing (NLP). This chapter begins by outlining the fundamental role of data annotation in NLP, highlighting its critical importance in training effective and accurate NLP models. Traditionally, this process has been labor-intensive and time-consuming, often requiring substantial human effort and resources. However, with the advent of advanced language models, there is a significant shift in how this task can be approached.

We delve deeply into how state-of-the-art LLMs, with their sophisticated understanding and processing of natural language, are revolutionizing the way we handle raw text data. These models are equipped to automatically annotate text data, a task that traditionally required extensive human intervention. The chapter discusses the mechanisms through which these LLMs operate, elaborating on their ability to interpret, analyze, and classify textual data with remarkable accuracy.

Further, the chapter investigates various methodologies employed in using LLMs for data annotation. This includes the use of these models to annotate unlabeled data sets as well as their ability to generate entirely new sets of labeled data. Such approaches are particularly beneficial in scenarios where labeled data is scarce or where the labeling process poses significant challenges.

An important aspect of this chapter is the discussion on the automation of the data labeling process. By integrating LLMs into this process, we have observed a substantial reduction in the time and resources traditionally required for data annotation. This automation does not compromise the quality of the output; in fact, it often enhances it, resulting in high degrees of label accuracy.

To validate these claims, we conducted a series of experiments focusing on four fundamental NLP tasks: Sentiment Analysis, Named Entity Recognition (NER), Relation Extraction (RE), and Aspect Sentiment Triplet Extraction (ASTE). These experiments were designed to test the efficiency and accuracy of LLMs in annotating data for these specific tasks. The results provide insightful data on the effectiveness of using LLMs in various NLP scenarios.

The chapter concludes by discussing the broader implications of our findings. The use of LLMs for data annotation not only streamlines the data preparation process in NLP but also significantly improves the quality of the training data. This, in turn, leads to the development of more robust and reliable NLP models. We emphasize that this advancement in data annotation methods has the potential to transform the landscape of NLP research and application, paving the way for more advanced, efficient, and accurate NLP systems in the future.

1.3 Thesis Outline

This thesis is meticulously designed to systematically delve into and significantly contribute to the realm of natural language processing (NLP) and machine learning, particularly focusing on environments where resources are scarce. The structure of the thesis is such that it facilitates a thorough exploration of these fields. Chapter 1, aptly titled "Introduction", sets the stage for the research by outlining the central research problem and the specific settings within which this problem is situated. This chapter serves as the gateway to the thesis, providing the necessary context and background for the ensuing discussions. In Chapter 2, "Literature Review", the thesis builds a strong foundation by meticulously examining the existing body of research. This chapter is crucial as it not only reviews prior work but also introduces key concepts and methodologies that are pivotal to the field of NLP and machine learning. The comprehensive review presented in this chapter ensures that the reader is well-equipped with the necessary background knowledge to understand the innovative approaches and experimental techniques that are discussed in the later chapters. Chapter 3, titled "A Generation-Based Data Augmentation Approach for Low-Resource Information Extraction," marks a shift towards more practical applications. In this chapter, an avant-garde technique is introduced, which leverages the power of generative models to synthesize diverse datasets. This approach is particularly beneficial for enhancing the performance of information extraction systems in scenarios where data is limited, a common challenge in low-resource settings. The focus then transitions in Chapter 4, "Developing a Task-Specific Multilingual Dialogue System Incorporating Local Entities and Contexts in a Global Setting." This chapter underscores the development of a sophisticated multilingual dialogue system. The novelty of this system lies in its ability to seamlessly integrate local specifics and nuances within a globally applicable framework, addressing a critical need in the field of NLP. In the penultimate chapter, Chapter 5, "Utilizing Advanced Language Models for Enhanced Data Labeling in Natural Language Processing Tasks," there is an exploration of how cutting-edge language models can revolutionize the data labeling process. Data labeling is a fundamental step in machine learning and AI development, and this chapter highlights how these advanced models can streamline and refine this process, thereby contributing significantly to the field. Finally, the thesis culminates with Chapter 6, "Conclusions and Future Directions." This chapter synthesizes the

findings of the thesis and offers a reflective overview of the entire research. It not only summarizes the key contributions made in each chapter but also casts a vision for future research directions in the field of Natural Language Processing. This conclusive chapter ensures that the thesis not only provides a comprehensive understanding of the current challenges and advancements in managing low-resource scenarios in AI and NLP but also sets the stage for future explorations in this dynamic and evolving field.

Below are the summary for the following chapters:

Chapter 2: Literature Review
In Chapter 2, we delve into the existing literature and foundational concepts crucial for understanding this thesis. We will explore various methods developed to enhance neural model performance in scenarios with limited resources. This chapter also includes an overview of pertinent tasks in natural language and image processing, setting the stage for the methodologies and experiments discussed in subsequent chapters.

Chapter 3: A Generation-Based Data Augmentation Approach for Low-Resource Information Extraction
Chapter 3 presents an innovative generation-based data augmentation technique designed to bolster information extraction in low-resource settings. This approach leverages advanced generative models to create diverse and representative datasets, addressing the challenges of data scarcity. We will demonstrate how this method significantly improves the accuracy and robustness of information extraction systems, especially in languages and domains where data is sparse.

Chapter 4: Developing a Task-Specific Multilingual Dialogue System Incorporating Local Entities and Contexts in a Global Setting
In Chapter 4, we focus on the development of a multilingual dialogue system tailored to specific tasks, integrating local entities and contexts within a global framework. This chapter details the challenges and solutions in designing systems that can effectively operate across multiple languages and cultural contexts. We will explore how incorporating local nuances enhances the system's performance and relevance in a globalized world.

Chapter 5: Utilizing Advanced Language Models for Enhanced Data Labeling in Natural Language Processing Tasks

Chapter 5 explores the use of cutting-edge language models to improve data labeling in various natural language processing tasks. We discuss the potential of these models to automate and refine the data annotation process, thereby reducing the time and resources required for manual labeling. This chapter highlights the effectiveness of these models in generating high-quality, reliable labels, especially in complex and nuanced linguistic scenarios.

Chapter 6: Conclusions and Future Directions
Chapter 6 integrates the various outcomes of this dissertation, presenting a contemplative summation of the overall investigation. This chapter accomplishes more than a mere recapitulation of the principal contributions identified in each segment; it also projects potential avenues for forthcoming inquiries within the realm of Natural Language Processing. Serving as the final chapter, it guarantees that the thesis delivers an exhaustive comprehension of both the existing obstacles and progressions in handling scenarios of limited resources in Artificial Intelligence and Natural Language Processing. Furthermore, it establishes a foundation for future investigative pursuits in this rapidly changing and developing domain.

Chapter 2

Literature Review

In this section, we elaborate on pertinent background information and foundational concepts that are crucial for grasping the subsequent segments of this thesis. Initially, we delve into prevalent methods designed to minimize overfitting, a common challenge in model training. Following this, our focus shifts to the exploration of semi-supervised learning and transfer learning. These approaches enhance model efficacy by utilizing additional resources, including unlabeled datasets and training materials from various domains or tasks. To conclude, this chapter provides an in-depth examination of related tasks in natural language processing and image processing, offering further insights into these critical areas.

2.1 Methods for Improving Generalization

2.1.1 Data Augmentation

Data Augmentation (DA) is a methodology employed in machine learning to enrich the diversity of training data sets without the necessity of acquiring additional data. This is achieved through techniques that either create variations of the existing data or synthesize new data. The overarching aim of DA is to function as a regularizing mechanism, contributing significantly to the reduction of overfitting during the training of machine learning models. As elucidated by Shorten and Khoshgoftaar in their seminal work [13], as well as by Hernández-García and König [24], this approach is particularly widespread in the field of computer vision (CV).

Within CV, common practices such as image cropping, flipping, and color adjustments are fundamental to the model training process. Conversely, in the realm of Natural Language Processing (NLP), where inputs are inherently more discrete, the development of effective DA strategies that preserve essential invariances poses a significantly more intricate challenge, as discussed by Feng et al. [25].

In the domain of NLP, despite numerous inherent challenges, a diverse array of DA techniques has been conceptualized and implemented. These techniques span from relatively straightforward rule-based manipulations, as noted by Zhang et al. [26], to more sophisticated generative approaches, as explored by Liu et al. [27]. The primary objective of DA in this context is to provide a viable alternative to the costly and time-consuming process of collecting new data. The ideal DA methodology should not only be straightforward to deploy but also significantly enhance the performance of the model, as indicated by Feng et al. [25]. Nevertheless, most DA methods necessitate a trade-off between ease of implementation and effectiveness. Rule-based approaches, while relatively simple to implement, often yield only modest enhancements in model performance, as identified by Wei and Zou [20]. In contrast, techniques that employ trained models, though potentially more complex and costly to implement, can introduce a wider spectrum of data variations, potentially leading to marked improvements in performance. Customizing model-based techniques to specific downstream tasks can substantially influence performance, yet such customization presents its own set of developmental and applicational challenges. Furthermore, it is crucial that the distribution of augmented data neither closely mirrors nor greatly diverges from that of the original data. An overly similar distribution can lead to overfitting, while a highly divergent one may result in the model being trained on unrepresentative examples, ultimately yielding suboptimal performance. As such, devising effective DA strategies requires a careful balance between maintaining similarity to and differing from the original data, as underscored by Wei and Zou [20].

2.1.2 Active Learning

Active Learning (AL) is a methodology that prioritizes the optimization of model performance while necessitating a minimal number of sample annotations [28]. This approach stands in stark contrast to Deep Learning (DL), a paradigm that

is inherently data-hungry. DL requires a substantial volume of data to effectively fine-tune its numerous parameters, thereby facilitating the extraction of enhanced features. The recent proliferation of internet technology has catapulted us into a data-rich epoch [29]. This deluge of data has significantly elevated the prominence of DL, triggering a wave of rapid technological progressions. In comparison, AL has not attracted as much attention as DL. This discrepancy can be traced back to the era preceding the ascendance of DL, where traditional machine learning methodologies did not demand a large number of labeled samples. Consequently, the early developments in AL were somewhat overlooked.

Despite the impressive achievements of DL in various fields, largely propelled by the availability of extensive, publicly accessible labeled datasets, the process of gathering large volumes of high-quality annotated data remains a laborious and often impractical task, especially in specialized domains such as speech recognition [30], information extraction [31], and medical imaging [32, 33]. In light of these challenges, the significance and utility of AL are increasingly being acknowledged. The ability of AL to function effectively with fewer annotated samples makes it particularly valuable in scenarios where acquiring annotated data is difficult or resource-intensive [34]. This recognition marks a crucial shift in the landscape of machine learning, highlighting the complementary roles of both AL and DL in advancing the field.

2.1.3 Knowledge Distillation

Knowledge Distillation (KD) [35] represents a significant advancement in the field of machine learning, particularly in the context of model efficiency and generalization. This technique, centered around the concept of a smaller "student" model learning from a more extensive "teacher" model, aims to retain the student model's manageability while enhancing its performance [36]. The following paragraphs delve into the intricacies of KD, encompassing its methodology, implications, and the challenges it presents.

At the core of KD lies the principle of model compression and efficiency [37, 38]. This method is instrumental in shrinking the model's size, where a less complex student model is trained under the guidance of a more sophisticated teacher model. The primary goal is to preserve the student model's compactness while minimizing

the compromise on its effectiveness. This approach is not only resource-efficient but also makes advanced models more accessible for applications with limited computational capabilities.

KD's knowledge transfer process is classified into three distinct levels, each contributing uniquely to the student model's learning [39]. Firstly, universal knowledge transfer occurs, often facilitated through label smoothing, acting as a regularizer. Secondly, domain-specific knowledge is imparted, wherein the student model gains insights into class relationships, notably at the logit layer. Lastly, individual instance knowledge is transferred, allowing the student model to tailor its learning to the intricacies of each specific event.

The mechanisms of knowledge representation and transfer in KD are diverse, mimicking various facets of the teacher model, such as its representation space, decision boundaries, or intra-data relationships [36]. Some advanced KD techniques even involve collaborative learning strategies among multiple student models, further enriching the learning process.

Despite its evident success, KD poses significant challenges and open questions [40]. One primary concern is identifying the exact locus of knowledge within neural networks and devising optimal strategies for its transfer from the teacher to the student model. This challenge underscores a fundamental gap in our understanding of neural network operations and learning processes.

KD has demonstrated its ability to enhance learning efficiency, especially in the face of challenges like label noise and class imbalance [41, 42]. It consistently outperforms standard training methods, offering generalization gains that are crucial for real-world dataset applications. However, the optimization of the student model and the nature of the dataset used for distillation significantly influence the effectiveness of KD. There is an interesting observation that a closer alignment with the teacher's predictive distributions does not necessarily translate to superior generalization in the student model.

In conclusion, Knowledge Distillation emerges as a potent tool in model compression and enhancing the generalization capabilities of smaller models [43]. It employs a multifaceted approach to knowledge transfer, spanning various levels of complexity. While KD offers considerable benefits in model efficiency and performance, it also presents unresolved challenges in optimization and knowledge transfer. These

challenges play a crucial role in determining the student model's capacity for effective generalization, highlighting the need for continued research and development in this domain.

2.1.4 Regularization and Weight Decay

Regularization and weight decay are fundamental concepts in the field of machine learning, particularly within the context of preventing overfitting in model training [44, 45]. Regularization refers to the process of adding information or constraints to a model to prevent it from fitting the noise in the training data. This is achieved by introducing a regularization term in the loss function, which penalizes complex models and thus promotes simpler solutions that may generalize better to new, unseen data [46].

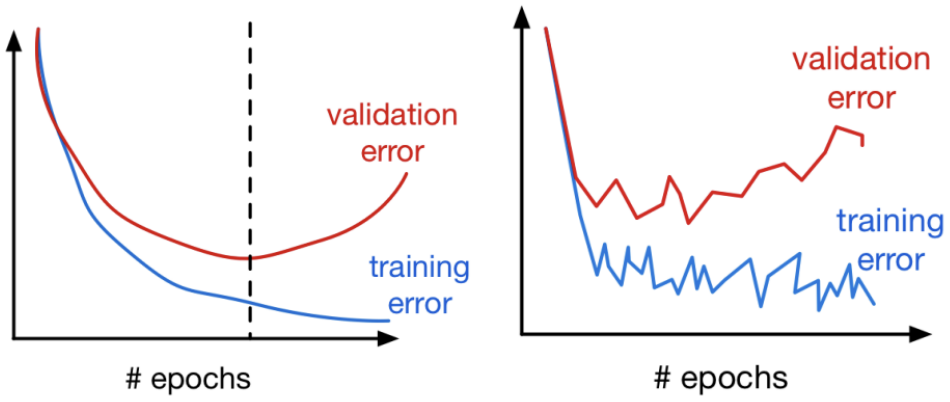


FIGURE 2.1: Training curves illustrate the correlation between the quantity of training iterations and both the training and test errors. The left side depicts an idealized model, while the right side takes into account variations in the error due to randomness in the stochastic gradient descent (SGD) updates.

Consider the training curves depicted in Figure 2.1, illustrating the correlation between the number of training iterations and the training and test error. The left side portrays an idealized version, while the right side takes into account error fluctuations due to the stochastic nature of the Stochastic Gradient Descent (SGD) updates.

Consequently, the overall cost can be expressed as $E(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y(x,\theta),t) + R(\theta)$, where the first part represents the training loss and the second part, the

regularizer. To illustrate, let's assume a linear regression model is being trained with two inputs, x1 and x2, which are identical in the training set. Although the weights in Figure 5 yield identical predictions on the training set and are equivalent in terms of minimizing the loss, Hypothesis A is preferable due to its anticipated stability under changing data distributions. For instance, with an input of (x1 = 1, x2 = 0) in the test set, Hypothesis A predicts 1, whereas Hypothesis B predicts -8, with the former being more plausible. Ideally, a regularizer should favor Hypothesis A by assigning it a lesser penalty.

One such regularizer is the L2 regularization, which, despite the terminology, corresponds to the squared L2 norm. For a linear model, this is defined as $R_{\mathrm{L2}}(\mathbf{w}) = \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2$. Here, λ, sometimes referred to as the weight cost, is a hyperparameter. L2 regularization tends to favor hypotheses with smaller weight norms. For example, with $\lambda = 1$, it assigns a penalty of 1 to Hypothesis A and 82 to Hypothesis B, significantly favoring Hypothesis A. This regularization approach, which includes both the training loss and the regularizer, motivates the training algorithm to seek a balance between fitting the training data and the weight norms. L2 regularization is easily extendable to neural networks by penalizing the sum of the squares of all the weights across all layers.

Weight decay, a specific form of regularization, operates by applying a penalty on the magnitude of the model's weights, typically in the form of an L2 norm [47]. This approach effectively constrains the weights, encouraging the model to distribute importance across features more evenly and to avoid relying excessively on any single feature. Both regularization and weight decay are critical in enhancing the robustness and generalizability of machine learning models, ensuring they perform well not just on the training dataset but also on new, external data.

Incorporating regularizers into stochastic gradient descent (SGD) computations is relatively straightforward. Due to the linearity of derivatives, the partial derivative of E with respect to θ_j can be broken down as shown in equations (4) to (8). Particularly in the context of L2 regularization, this leads to an insightful interpretation in SGD, where the weights are reduced by a factor of $1 - \alpha\lambda$ in each iteration, giving rise to the term "weight decay."

Regularization stands as a cornerstone concept in machine learning, with numerous theoretical justifications [48]. Regularizers are often seen as penalizing a network's

"complexity" or favoring more "probable" explanations. These perspectives can be formalized in certain idealized scenarios. However, they are challenging to precisely define in the context of neural networks and fail to explain many observed practical phenomena. Therefore, this discussion will not extend beyond the aforementioned justification for weight decay.

2.1.5 Stochastic regularization

In recent years, a significant advancement in the field of neural networks has been the incorporation of stochastic elements to enhance generalization capabilities. Unlike the deterministic function computation in traditional network architectures, the introduction of stochasticity can mitigate pathological behaviors and reduce the tendency for overfitting [49]. This approach is commonly referred to as stochastic regularization. Although this term does not imply the addition of a regularization term to the cost function, it serves a similar purpose.

The most prevalent method of stochastic regularization is known as dropout [50]. The mechanism of dropout is straightforward: it involves deactivating each unit in the network with a certain probability ρ (often set at 0.5), effectively setting its activation to zero. This process is mathematically represented by multiplying the activations with a binary mask variable m_i, which randomly assumes values of 0 or 1, as shown in the equation $h_i = m_i \cdot \phi(z^{(i)})$. The backpropagation equations are derived conventionally, taking into account the effect of the mask variable on the activations.

Dropout is particularly effective because it biases the network towards certain configurations over others, similar to the effect of L2 regularization. For instance, when inputs have a 50% chance of being dropped out, configurations that would otherwise exhibit high variance in their predictions (and consequently, higher error rates on the training set) are less favored. This similarity with L2 regularization's objectives highlights the utility of dropout.

It is crucial, however, to manage the stochasticity during the testing phase to avoid unpredictable outcomes. A naïve solution of excluding dropout during testing can lead to discrepancies, as the units would be exposed to double the amount of incoming signals compared to the training phase. To counter this, the weight

values are adjusted by a factor of $1 - \rho$ during testing, ensuring consistency in the network's responses.

Dropout has rapidly become an essential component in neural network training, often resulting in considerable performance improvements even when combined with other techniques [51]. Additionally, other forms of stochastic regularization, such as batch normalization, have been introduced. While batch normalization is primarily associated with optimization, it also offers regularization benefits. Interestingly, the inherent stochasticity in stochastic gradient descent, typically viewed as a limitation, can act as a form of regularization. Although the intricacies of stochastic regularization are not fully understood, its continued utility in neural network optimization is highly anticipated.

2.1.6 Ensemble Methods

Ensemble methods in machine learning, specifically in the context of neural networks, offer a robust approach to improve prediction accuracy [52]. The concept hinges on the idea that combining the predictions of multiple networks trained independently can significantly reduce the variance in predictions, leading to a lower overall loss [53]. This strategy effectively addresses the limitations inherent in using a single model.

However, the ideal scenario of training each network on a separate, independently sampled training set is often impractical. The alternative is to introduce variability into the training process of multiple networks using the same dataset. This can be achieved through several methods:

1. **Bagging**: Training each network on random subsets of the full training dataset. This approach ensures that each network learns from a slightly different perspective of the data, simulating the effect of independent training sets.

2. **Diverse Architectures**: Utilizing networks with varying structures—different numbers of layers, units, or activation functions. This diversity ensures that each network captures different aspects or patterns in the data.

3. **Different Models or Algorithms**: Beyond varying architectures within a similar model framework, employing entirely different models or learning algorithms can introduce a higher level of variability.

The collection of models used for prediction is known as an "ensemble." The effectiveness of ensembles is well-documented, as they often outperform individual networks. This superiority is not just theoretical; it's empirically observed in many machine learning competitions, such as ImageNet and Netflix challenges, where ensembles are frequently the key to winning entries [54].

The theoretical underpinning for the superiority of ensembles, especially in cases of convex loss functions, can be explained using a mathematical framework. If the loss function L is convex with respect to the output predictions y, then the loss of the ensemble's averaged prediction is mathematically guaranteed to be lower than the average losses of individual predictions. This is a direct consequence of the definition of convexity:

$$L(\lambda_1 y_1 + \cdots + \lambda_N y_N, t) \leq \lambda_1 L(y_1, t) + \cdots + \lambda_N L(y_N, t)$$

for $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.

It's important to note, however, that this guarantee does not necessarily hold for non-convex loss functions, such as error rates. Despite this, ensembles are still highly effective in practice for a broad range of applications, offering a pragmatic solution in scenarios where the loss function is not convex. The principle behind this effectiveness is related to the Rao-Blackwell theorem from statistics, which suggests that averaging over a set of estimators can lead to a more accurate overall estimate.

In summary, ensembles represent a powerful strategy in machine learning, leveraging the strength of multiple models to achieve greater predictive accuracy and robustness than individual models could achieve alone [55].

2.2 Transfer Learning

In the modern era, the process of allocating resources for data acquisition has become increasingly arduous. This difficulty arises from the scarcity of data, its inaccessibility, the expense involved in obtaining it, and the complexity inherent in its compilation. In response to these challenges, there has been a shift towards exploring alternative methodologies for data collection. A particularly notable strategy in this context is the concept of transferring knowledge between disparate tasks. This strategy has been pivotal in fostering the growth of Transfer Learning (TL), a paradigm specifically designed to refine the processes of data gathering and learning in the field of machine learning (ML) [56]. Typically, ML algorithms are engineered with the aim of predicting future occurrences, with a primary focus on addressing specific tasks independently. Transfer Learning, however, deviates from this conventional model by incorporating data from a 'source task' and applying it to a 'target task'. This method holds the potential for generating more effective solutions [57].

The fundamental goal of TL is to enhance the understanding of a current task by linking it to other tasks that may have been performed at different times, albeit within a similar domain [58]. This linkage is critical in bolstering the learning process, as it establishes a connection between the source task and the target task, leading to more logical, expedient, and improved solutions. TL is designed to provide a more effective mode of learning and interaction between the source and target tasks, introducing a dynamic aspect of debate in the learning trajectory [59]. Additionally, TL is particularly advantageous in scenarios where the target task suffers from a scarcity of training data. The strategic significance of TL extends beyond its utility in the immediate task at hand, encompassing its applicability across a variety of tasks. However, it is vital to recognize that the interaction between the source and target tasks is not always beneficial. In instances where the transference of test and training samples results in a deterioration of the target task's performance, this phenomenon is labeled as 'negative transfer'. Conversely, when such transference leads to improved performance, it is known as a 'positive transfer'.

2.2.1 Cross Lingual Transfer

The aim of cross-lingual transfer lies in leveraging the extensive knowledge base of high-resourced languages, such as English, to aid languages with limited / low resources [60]. This endeavor is significantly facilitated by the use of both cross-lingual contextual and static word embeddings, leading to commendable outcomes in a variety of transfer tasks [61]. The technique of mapping static embeddings, which is thoroughly investigated, hinges on the hypothesis that the embedding spaces of different languages possess analogous structures. This approach includes the alignment of these spaces via an orthogonal matrix, culminating in a unified embedding space. Further advancements in this domain have seen the expansion of this projection technique to encompass the alignment of contextual representations as well [62].

However, the foundational assumption of isomorphism underlying this approach has encountered skepticism, prompting an inquiry into alternative mapping methodologies, specifically those employing non-linear techniques [63]. Another significant strand of research in this field is joint training. Initially, this research direction concentrated on the simultaneous acquisition of static word embeddings for multiple languages. More recently, the focus has shifted towards the pretraining of cross- or multi-lingual language models. Within the unsupervised domain, multilingual language models have emerged, demonstrating noteworthy efficacy in the realm of cross-lingual transfer. This evolution signifies a continuous exploration and enhancement in the methodologies employed for cross-lingual linguistic resource sharing and development.

The evolution of cross-lingual transfer techniques has been marked by a growing emphasis on contextualized language models. These models, unlike static embeddings, capture the context in which words appear, leading to more nuanced and accurate representations of language. This shift towards contextualized models reflects the broader trend in natural language processing towards deep learning-based approaches. The efficacy of these models in cross-lingual transfer is evident in their ability to handle subtleties and nuances in languages, which are often lost in static models.

In addition to these technical advancements, there has been a growing recognition of the importance of linguistic diversity in machine learning models. This has

led to an increased focus on low-resource languages, which have traditionally been underrepresented in language technology. The drive to include these languages is not only a matter of fairness and representation but also enhances the robustness and versatility of the models themselves. By training models on a diverse range of languages, researchers are able to uncover and address biases and limitations that might not be apparent when focusing solely on high-resource languages.

The intersection of cross-lingual transfer with other areas of machine learning, such as transfer learning and domain adaptation, is also a rich area of exploration [64]. Transfer learning techniques, which involve applying knowledge gained in one domain to a different but related domain, have proven particularly useful in cross-lingual settings. These techniques allow for the leveraging of large datasets available in high-resource languages to improve performance in low-resource languages. Similarly, domain adaptation techniques help in adjusting models trained on data from one domain (e.g., news articles) to perform well on data from a different domain (e.g., medical texts) [65].

Moreover, there is an increasing interest in the ethical implications of cross-lingual transfer. Questions about representation, bias, and the potential perpetuation of inequalities through language technologies are gaining prominence. Researchers are beginning to address these issues by developing more inclusive methodologies and striving for transparency in their models. This includes the creation of guidelines for ethical research practices in cross-lingual NLP and efforts to ensure that the benefits of language technology are equitably distributed across different linguistic communities [66].

In conclusion, the field of cross-lingual transfer is undergoing rapid and significant transformations. The move towards more sophisticated, context-aware models, coupled with a heightened awareness of ethical considerations and the push for linguistic diversity, is shaping the future of language technology. As the field continues to evolve, it promises to bring forth innovations that not only enhance the capabilities of machine learning models but also contribute to a more inclusive and equitable technological landscape.

2.2.2 Cross Domain Transfer

In earlier discussions, we delved into the concept of cross-domain transfer, an integral component of the transductive approach [67]. This approach is widely embraced in the realms of Natural Language Processing (NLP) and Computer Vision (CV), where it plays a pivotal role in task execution. A quintessential illustration of cross-domain transfer can be observed in the realm of sentiment analysis. Insights derived from analyzing sentiments in movie reviews can be adeptly applied to the evaluation of book reviews. This exemplifies the seamless application of knowledge across different but related domains.

Another notable instance of cross-domain transfer is evident in the adaptation of models originally crafted for the ImageNet database. These models are retooled for specialized image classification tasks, demonstrating the versatility and adaptability of cross-domain methodologies. In such instances, it is a common assumption that the label sets for both the source (original) and target (new) domains remain consistent. This assumption forms the bedrock upon which strategies are developed to extract and utilize domain-independent knowledge from the source domain. Such strategies are instrumental in augmenting the effectiveness of the model when applied to the target domain.

The concept of cross-domain transfer shares parallels with cross-lingual transfer. In both cases, the overarching goal is to enhance performance in a new domain by leveraging insights and methodologies from a different but related domain. To achieve this, various techniques are employed. These include the alignment of features, which ensures consistency and relevance of the features used in both domains. Shared parameterization is another technique where parameters used in models for one domain are applied to another, ensuring a certain level of uniformity and efficiency in model performance. Finally, the weighting of instances or domains plays a crucial role. This involves assigning different levels of importance or priority to certain aspects or areas within the domains, thereby fine-tuning the model's focus and effectiveness.

However, it's important to acknowledge the underlying assumptions and challenges in cross-domain transfer. One such assumption is the consistency of label sets across the source and target domains. While this assumption simplifies the transfer process, it also imposes constraints on the applicability of transfer learning in

scenarios where label sets differ significantly. This calls for innovative approaches to manage such discrepancies and ensure effective transfer.

Moreover, the strategies employed in cross-domain transfer, such as feature alignment [68], shared parameterization [69], and instance or domain weighting [70], are not just technical maneuvers but also represent a deeper understanding of the nature of knowledge transfer [71]. Feature alignment, for instance, is about identifying and leveraging commonalities in data representation across domains. Shared parameterization speaks to the universality of certain model parameters, while instance weighting addresses the nuanced differences between domains and how to prioritize them.

These strategies are integral to the success of cross-domain transfer and highlight the nuanced understanding required to implement this approach effectively. They are not just about technical adaptation but also about conceptual adaptation - understanding the core principles that govern data and models in different domains and how these can be harmonically aligned for optimal performance.

The parallels between cross-domain and cross-lingual transfer further underscore the universal applicability of these concepts [72]. Just as in cross-domain transfer, cross-lingual transfer involves the application of knowledge from one language to another. This is evident in tasks such as machine translation, where models trained on one language pair are adapted for another. The underlying principles remain the same - leveraging existing knowledge and methodologies to achieve efficiency and effectiveness in a new domain.

In conclusion, cross-domain transfer is a cornerstone of modern AI, pivotal in driving forward the fields of NLP and CV among others [73]. Its relevance extends beyond mere technical application; it represents a paradigm shift in how we approach problem-solving in AI. By embracing the principles of adaptability, scalability, and efficiency inherent in cross-domain transfer, we can continue to push the boundaries of what AI can achieve, making it more versatile, effective, and accessible across various domains and applications.

2.3 Zero-shot and Few-shot Learning

2.3.1 Zero-shot Learning

Zero-shot learning is a method that employs a collection of labeled training instances within a particular feature space [74]; these instances are identified as the seen classes [75]. Concurrently, within the same feature space, there is a set of unlabeled test instances [76]. These instances correspond to a distinct set of classes, which are referred to as the unseen classes. The nature of the feature space in zero-shot learning is typically constituted by a space of real numbers. In this context, each instance, whether labeled or unlabeled, is depicted as a vector situated within this numerical space. A fundamental assumption in zero-shot learning is the notion that each instance is exclusively associated with a single class, negating the possibility of multi-class affiliation for any given instance [77].

Definition 2.3.1 (Zero-Shot Learning)**.** Let $\mathcal{D}\text{tr}$ be a set of labeled training instances corresponding to seen classes $\mathcal{S}$. Zero-shot learning refers to the process of devising a classifier $f_u(\cdot) : \mathcal{X} \to \mathcal{U}$, which is trained on $\mathcal{D}\text{tr}$ and capable of classifying test instances $\mathcal{X}\text{te}$ (i.e., predicting $\mathcal{Y}\text{te}$) that belong to the unseen classes $\mathcal{U}$.

Based on the provided definition, we understand that zero-shot learning primarily focuses on utilizing the knowledge from training instances, denoted as D_t^r, for the purpose of classifying test instances. This approach is distinct because it involves non-overlapping label spaces in training and testing phases. Essentially, zero-shot learning is a specialized branch of transfer learning, as referenced in [114, 115]. Transfer learning involves transferring knowledge from a source domain and task to a target domain to develop a model for the target task, as mentioned in [114, 115].

As per sources [23, 114], transfer learning can be divided into two types: homogeneous and heterogeneous, depending on the similarity of feature and label spaces in the source and target domains/tasks. Homogeneous transfer learning maintains the same feature and label spaces, whereas heterogeneous transfer learning deals with different feature and/or label spaces. In the context of zero-shot learning, the feature space for both training (source) and testing (target) instances is identical

and denoted as X. However, the label spaces differ: the source label space comprises the seen class set S, and the target label space includes the unseen class set U. Therefore, zero-shot learning falls under the category of heterogeneous transfer learning, specifically a type that features different label spaces, abbreviated as HTL-DLS [78].

Numerous methods have been developed for HTL-DLS, often under scenarios where some labeled instances are available for the target label classes, as discussed in [79]. However, zero-shot learning is unique in that it lacks labeled instances for the target label space classes (unseen classes), setting it apart from other problems in HTL-DLS [80].

2.3.2 Few-shot Learning

Few-Shot Learning (FSL) draws inspiration from the remarkable logical and analytical capabilities of humans and is frequently utilized in edge computing scenarios [81]. This concept revolves around the idea of a computer program enhancing its effectiveness in specific tasks and measurements through a modest amount of learning experiences. Notably, these learning experiences in FSL are quite limited [82]. This approach mirrors how human cognitive skills relate to distinct memory systems as highlighted by current neuroscientific studies. These systems include the gradual learning process in the neocortex and the rapid learning mechanism in the hippocampus, which correlate with FSL's slow, data-oriented learning and quick, feature-focused learning [83].

To fully comprehend FSL, it's essential to understand two principal concepts: the N-way-K-shot dilemma and cross-domain FSL [84]. The N-way-K-shot issue is a fundamental challenge in FSL, characterized by a limited training dataset, known as the support set, that serves as a reference during a testing phase. The query set, where predictions are made, encompasses categories not encountered in the support set. In a standard N-way-K-shot setting, the support set includes N categories, each with K instances, making up a total of N*K instances. Thus, N-way-1-shot refers to one-shot learning, while N-way-0-shot indicates zero-shot learning. Cross-domain FSL, evolving from the principles of transfer learning, entails applying insights

gained in one domain to another, often overcoming domain disparities. This fusion of cross-domain elements with FSL principles marks a recent and complex advancement in this area.

Due to restricted access to comprehensive datasets, Few-Shot Learning (FSL) is typically dependent on a limited selection of examples [85]. This dependence often results in biased estimations that don't accurately represent the full spectrum of data. This issue becomes particularly critical as it can negatively impact the accuracy of various tasks. To mitigate these challenges, several strategies are essential. These include enhancing data through data augmentation techniques, delving into the details of features within and between classes, and creating tailored images / text to better approximate true data distribution [86].

Below are some key challenges and research problems in FSL:

- **Sensitivity to Feature Reutilization**: This aspect of FSL emphasizes the accumulation of knowledge by leveraging extensive auxiliary datasets [87]. The methodology employs transfer learning techniques, wherein knowledge from a primary, well-established domain is repurposed for a related, yet distinct target domain. This process typically involves pre-training for extracting complex, high-dimensional features, followed by fine-tuning to make minor adjustments to the initial parameters. Although this strategy is effective in enhancing the performance of specific tasks, it may encounter difficulties in generalizing across diverse tasks. Moreover, there is a potential risk of negative knowledge transfer, particularly when there is a significant discrepancy between the source and target domains.

- **Adaptability to Future Tasks**: Contrasting with transfer learning, meta-learning in the context of FSL focuses on rapid adaptation from familiar to novel tasks [88]. This approach involves the summarization of meta-knowledge across a spectrum of tasks, thereby facilitating the efficient integration of future tasks. While meta-learning provides a comprehensive framework for learning, its efficacy is somewhat constrained to scenarios where the training and testing tasks bear resemblance. Challenges arise from its inflexibility and the dependency on the underlying network architecture, which can limit its applicability in diverse learning environments.

- **Limitations of Single-Modal Information in FSL**: The dependence on single-modal information in FSL presents obstacles in effective feature learning due to the inherent limitations in the information available [89]. Enriching FSL with additional modalities, such as semantic assistance, can significantly augment the learning process. This approach involves the introduction or generation of semantic information, which serves as weak supervision. It enables an adaptive classification strategy that works in conjunction with the primary model, thereby enhancing the model's ability to learn from limited data through a more holistic understanding of the features and their relationships.

2.4 Pre-trained Langauge Models & Large Language Models

2.4.1 Pre-trained Langauge Models

Since their initial development, Pre-trained Language Models (PLMs) [90] have undergone substantial evolution, marking a significant shift in the field of Natural Language Processing (NLP). This evolution began with the inception of ELMo, a pioneering model that deviated from conventional static word representations by introducing context-sensitive word representations [91]. Unlike earlier models that treated words as fixed entities, ELMo employed a bidirectional Long Short-Term Memory (LSTM) network, adding depth to the understanding of language context. This model was distinguished by its two-phase training process: an initial pre-training phase, followed by a task-specific fine-tuning phase [92].

Building upon ELMo's foundation, the introduction of the Transformer architecture marked a turning point in the field. The Transformer is celebrated for its efficient parallel processing capabilities and innovative self-attention mechanisms, which significantly enhance the model's ability to understand language context. This architecture set the stage for the development of BERT (Bidirectional Encoder Representations from Transformers) [93]. BERT represented a leap forward in the field by employing bidirectional training, which allowed for a more nuanced understanding of language context. It achieved this by pre-training on a diverse

range of large, unlabeled datasets using unique tasks designed to capture the intricacies of language.

These advancements in context-aware word representations have substantially bolstered the performance of NLP tasks, leading to unprecedented improvements. This progress has not only redefined the capabilities of language models but has also ignited a wave of research in this domain. The "pre-training and fine-tuning" methodology, introduced by these models, has become a widely accepted and standard paradigm within NLP. In this approach, a PLM is first pre-trained on a large corpus of data to understand language at a general level and then fine-tuned for specific tasks, allowing for tailored applications across a spectrum of linguistic challenges.

Following this trend, a multitude of studies have emerged, exploring various architectural innovations and refinements. Models like GPT-2 [94] and BART [95] are notable examples of this ongoing exploration, each contributing unique elements to the evolving landscape of PLMs. These models have continued the trend of enhancing and refining the pre-training methods, further solidifying the significance of this approach. In essence, the paradigm of adapting a PLM to a wide range of downstream tasks involves a meticulous process of fine-tuning, which is pivotal in customizing the model's capabilities to specific linguistic requirements and applications.

2.4.2 Large Language Models

Recent studies in the field of artificial intelligence have unveiled a pivotal trend: the augmentation of the data capacity in Pre-trained Language Models (PLMs) markedly improves their efficacy across a diverse range of tasks. This trend is exemplified by models like GPT-3 [96], which comprises an astounding 175 billion parameters, and PaLM, which boasts an even more impressive 540 billion parameters. These findings are in harmony with the principles of the scaling law, a theoretical framework that suggests a direct correlation between the size of a model and its performance capabilities [97].

In contrast to their more modestly sized predecessors, such as the 330 million-parameter BERT and the 1.5 billion-parameter GPT-2, these expansive models

display a range of unique behaviors and capabilities. Notably, they exhibit emergent abilities in managing complex tasks that were previously unattainable by smaller models [98]. A prime example of this is the proficiency of GPT-3 in executing few-shot tasks through in-context learning, a feat that is notably challenging for its predecessor, GPT-2 [96].

This leap in performance and capability has led to the coining of the term "large language models (LLMs)" to describe these high-capacity PLMs. This nomenclature underscores their distinction from smaller models and highlights their advanced abilities. The interest in these LLMs has surged within the research community, given their transformative potential in various applications.

One of the most notable implementations of LLMs is seen in ChatGPT, which utilizes the GPT series' LLMs specifically for dialogic interactions [99]. ChatGPT stands as a testament to the advanced conversational abilities of these models, demonstrating an ability to engage in human-like dialogues with a level of sophistication and nuance that was previously unachievable in earlier iterations of language models. This advancement not only marks a significant milestone in the field of natural language processing but also opens new avenues for research and application in human-computer interaction.

2.5 Related Tasks and Techniques

2.5.1 Named Entity Recognition (NER)

Named entities are specialized phrases within a text that distinctly mark the names of individuals, organizations, geographical locations, among others [100]. To illustrate, consider the sentence, "U.N. official Ekeus is en route to Baghdad." This statement highlights named entities such as 'U.N.' and 'Ekeus.' The identification and systematic classification of these entities into predefined groups form a crucial aspect of extracting information from texts, a process known as Named Entity Recognition (NER). NER plays a pivotal role in structuring unstructured data by discerning and categorizing essential elements of information.

This task, however, presents significant challenges, primarily due to two reasons [101]. First, there is a notable lack of sufficiently annotated datasets for NER

purposes across numerous languages and domains. Such datasets are essential for training and evaluating NER systems, but their scarcity hampers progress in this field. Second, the restricted scope and variability of the available data pose a substantial barrier to achieving extensive generalization. This limitation is particularly evident in the constraints surrounding the diversity of words that can be recognized as named entities. These challenges necessitate innovative approaches in NER research, focusing on enhancing data availability and expanding the range of identifiable entities for more effective information extraction and analysis.

2.5.2 Part-of-Speech (POS) Tagging

Part-of-speech (POS) tagging is a critical process in computational linguistics where each word within a sentence is assigned to a specific grammatical category [102]. This task is a cornerstone for numerous operations in natural language processing (NLP) systems. It plays a pivotal role in enhancing the efficiency and accuracy of various NLP tasks, such as syntactic parsing, which involves analyzing the syntax of a sentence, and sentiment analysis, which focuses on determining the emotional tone behind a body of text.

In recent times, the field has witnessed significant advancements in POS tagging technologies [103]. Contemporary POS taggers are known for their high levels of accuracy. For instance, in well-established benchmarks like the PTB-WSJ, these taggers demonstrate an accuracy exceeding 97.96%, showcasing their robustness in understanding and categorizing linguistic elements [104]. Furthermore, when evaluated on a diverse array of 21 languages that are rich in linguistic resources, as part of the UD 1.2 dataset, these systems maintain an impressive average accuracy rate of over 96.50% [105]. Such figures are indicative of the substantial progress in the field, highlighting the sophisticated nature of current POS tagging systems.

Despite these achievements, there remain notable challenges in the realm of POS tagging. One significant issue is the variable performance of these systems across different languages, especially those that are less-resourced. The accuracy of POS taggers tends to decline in dealing with languages that have fewer linguistic resources available for training and analysis. Additionally, these systems often struggle with infrequent or rare words, reflecting a limitation in their ability to generalize

from training data to real-world applications. This variation in performance underscores the ongoing need for research and development in the field, aiming to enhance the adaptability and inclusiveness of POS tagging systems across diverse linguistic contexts.

2.5.3 Sentiment Analysis

Sentiment Analysis, a subfield of Natural Language Processing (NLP), involves the systematic identification and categorization of the sentiment polarity embedded within a text [106]. This process involves analyzing textual material, such as tweets, and classifying them into distinct sentiment categories: positive, negative, or neutral. This classification is based on the emotional tone conveyed in the text [107]. Leveraging both the textual data and its corresponding sentiment labels, machine learning models can be adeptly trained to accurately predict the underlying sentiment of the text.

The methodologies applied in Sentiment Analysis are diverse, encompassing several approaches. These include machine learning-based techniques, which rely on algorithmic models to learn from data [108]; lexicon-based strategies, which utilize a predefined list of words associated with specific sentiments [109]; and hybrid methods [110], which combine elements of both machine learning and lexicon-based approaches. Within the realm of Sentiment Analysis research, various specialized subcategories have emerged. These include multimodal sentiment analysis, which integrates multiple types of data such as text and images; aspect-based sentiment analysis, focusing on specific aspects or features within the text; fine-grained opinion analysis, which seeks to discern more nuanced sentiment categories; and language-specific sentiment analysis, tailored to specific linguistic contexts.

In recent advancements, deep learning techniques, notably models like RoBERTa [111] and T5 [112], have been instrumental in enhancing the performance of sentiment classifiers. These advanced models are trained using large datasets and are evaluated based on metrics such as F1-score, recall, and precision, which are crucial in determining the efficacy of the classifiers. To benchmark and evaluate these sentiment analysis systems, a variety of datasets are utilized. Notable among these are the Stanford Sentiment Treebank (SST) [113], the General Language Understanding Evaluation (GLUE) benchmark [114], and the IMDB movie review dataset

[115]. These datasets provide a comprehensive framework for assessing the performance of sentiment analysis models, ensuring their applicability and accuracy in real-world scenarios.

Aspect-based Sentiment Analysis (ABSA) Traditional studies in sentiment analysis have predominantly focused on determining the overall sentiment within sentences or entire documents, as evidenced by references in the literature. This approach usually operates under the assumption that a text exhibits a consistent sentiment towards a singular topic. However, such an assumption often proves to be inaccurate in the complexities of real-world contexts. In response to this limitation, the last decade has witnessed an increasing shift towards Aspect-Based Sentiment Analysis (ABSA) [107]. ABSA diverges from traditional sentiment analysis by focusing on the sentiments directed towards specific entities or attributes within these entities, rather than generalizing across the whole text. Within the scope of this discussion, an entity can encompass a specific product in the E-commerce sector, with attributes such as price and size [116]. Here, an entity can also be conceptualized as a broader aspect. Thus, in our discourse, we refer to both entities and their attributes under the umbrella term "aspects." The aim of ABSA is to provide intricate opinion summaries at this aspect level, offering in-depth sentiment insights crucial for a variety of downstream applications [117].

ABSA synthesizes a range of detailed sentiment analysis tasks, with the aspect target being the focal point [118]. For example, in a hypothetical scenario, the aspect targets could be "Windows 8" and "touchscreen functions". A fundamental component of ABSA is Aspect Sentiment Classification (ASC), which entails identifying the sentiment polarity associated with a specific aspect target [119]. However, it is not always possible to have a pre-determined aspect target. In such instances, Aspect Term Extraction (ATE) plays a critical role in discovering these targets [120]. Concurrently, Opinion Term Extraction (OTE) is instrumental in pinpointing opinion terms that significantly shape the sentiment polarity of a sentence or a specific target term [121]. The most intricate component of ABSA, Aspect Sentiment Triplet Extraction (ASTE), endeavors to offer a holistic view of sentiment information [122]. It does this by amalgamating an aspect target term, its related opinion term, and the sentiment expressed, into a cohesive triplet. An illustrative example of this could be the triplets: ("Windows 8", "not enjoy", Negative) and ("touchscreen functions", "not enjoy", Negative).

2.5.4 Task-oriented Dialogue System

Task-oriented dialogue systems represent a sophisticated integration of technology designed to streamline a multitude of tasks through interactive language-based communication [123]. These systems shine particularly in executing functions such as arranging ticket purchases, booking dining experiences, and offering aid in customer service scenarios [21]. At the heart of their operation lies a critical ability to discern and consistently monitor a user's specific needs during an exchange [124]. This is accomplished via a mechanism termed Dialogue State Tracking (DST) [125]. DST plays a pivotal role in identifying and preserving the user's requirements, articulated as slot-value pairs. Each of these pairs is a marker of an individual user need, ensuring that the system's replies are not only pertinent but also coherent.

The seamless functioning of these dialogue systems is anchored in four main components [21, 126]. First, there is Natural Language Understanding (NLU), which is tasked with interpreting the user's oral or textual inputs. Following this, the Dialogue State Tracking comes into play, keeping a detailed log of the ongoing conversation and the user's articulated needs. The third component, Dialogue Policy Learning, is instrumental in deciding the system's subsequent course of action, taking into account the present state of the dialogue. Finally, Natural Language Generation (NLG) comes into the picture, which is responsible for crafting the system's spoken or written outputs.

Together, these components form a synergistic framework, enabling task-oriented dialogue systems to emerge as indispensable tools in the realm of simplifying and automating interactions across various sectors. Their ability to handle complex tasks through intuitive language-based communication makes them a significant asset in enhancing user experience and efficiency in several domains.

Dialogue State Tracking The concept of Dialogue State refers to a comprehensive summary of a dialogue's progression up to a specific point, denoted as turn t [127]. This summary is crucial as it equips the conversational system with all the necessary information required to determine its subsequent course of action. Essentially, the Dialogue State is a reflection of the user's objectives within the conversation, represented through a series of (slot, value) pairings. These slots,

which are typically dependent on the specific domain of the conversation and predefined in the Ontology O, capture the goals or intentions of the user. Each slot is assigned a value that the user specifies to convey their conversational goal. For instance, in a dialogue pertaining to restaurant reservations, the Dialogue State at a particular turn might include pairs such as (FOOD, ITALIAN) and (AREA, CENTRE), indicating the user's preferences in terms of cuisine and location.

These slots fall into two categories: informable and requestable [125]. Informable slots encompass attributes that users can specify during the conversation to establish constraints, whereas requestable slots consist of attributes that users might inquire about from the system. In the context of restaurant reservations, for example, slots like FOOD, AREA, and PRICE would be informable, allowing users to set their preferences, while PHONE and ADDRESS would be requestable, providing users with information upon request.

The process of tracking these Dialogue States is crucial for understanding the progression of the conversation. This tracking involves identifying and maintaining the user's goals at each turn.

Dialogue State Tracking (DST) plays a vital role in this process. Its primary function is to estimate the current Dialogue State by predicting the relevant slot-value pairs at each turn. This prediction process can be executed in two distinct ways: turn-level prediction and dialogue-level prediction [124].

Turn-level prediction involves an update mechanism that can either be rule-based or learned. In the rule-based approach, the system makes predictions for the current turn and then integrates these with the previous Dialogue State using predefined rules. These rules might range from simple overwriting of values to more complex probabilistic combinations. Alternatively, in the learned approach, a function is trained to emulate the update process, taking into account both the previous Dialogue State and the current turn's predictions to forecast the current Dialogue State. This approach can be realized either through a dual-component system or a unified end-to-end model [128].

On the other hand, dialogue-level prediction entails the model considering the entire dialogue history at each turn to predict the complete Dialogue State. While this method provides a comprehensive view of the current state, it may lead to

inconsistencies as it does not take previous Dialogue States into account, potentially leading to discrepancies between consecutive states.

Overall, the management and prediction of Dialogue States are fundamental to the efficacy of conversational systems, ensuring they are responsive and relevant to user inputs throughout the dialogue [129].

2.5.5 Relation Extraction

Relation Extraction (RE) constitutes a critical process in the field of computational linguistics, where the primary objective is to deduce and categorize the relationships and characteristics associated with entities within a given sentence [130]. To illustrate, consider the sentence "Barack Obama was born in Honolulu, Hawaii." In this context, the goal of a relation classifier is to accurately determine and assign the specific relation, in this case, "bornInCity". This extraction process is not merely a standalone task but serves as a foundational element in the construction of relation knowledge graphs. Its significance extends deeply into various applications of natural language processing (NLP), encompassing a wide range of functionalities including, but not limited to, structured search queries, sentiment analysis, question-answering systems, and the generation of concise summaries. These applications demonstrate the profound impact and necessity of effective Relation Extraction in advancing the capabilities and understanding within the domain of NLP.

The importance of relation extraction goes beyond mere categorization [131]. It acts as a cornerstone in the development of relational knowledge graphs. These knowledge graphs are essential for visualizing and structurally representing the relationships between various entities. By mapping out these relationships in a graph format, it becomes easier for machines to understand and interpret complex data, leading to more accurate and efficient processing of large volumes of text.

In the broader scope of Natural Language Processing (NLP), relation extraction has profound implications [132]. Its applications are diverse and encompass several key areas. For instance, in structured search queries, relation extraction allows for more precise and context-aware search results [133]. In sentiment analysis, understanding the relationships between entities can provide deeper insights into the

sentiments expressed in the text. Similarly, in question-answering systems, relation extraction helps in accurately understanding the query and fetching the relevant information by understanding the relationships between different entities mentioned in the knowledge base. Furthermore, in the generation of concise summaries, relation extraction plays a pivotal role by identifying the most critical relationships and entities, which need to be included in the summary for it to be coherent and informative [134].

Additionally, the advancements in relation extraction have a significant impact on the development of intelligent systems capable of understanding and interacting with human language in a more sophisticated manner. These systems can be used in various applications, such as virtual assistants, automated customer support, and intelligent tutoring systems, where understanding the relationships between entities is crucial for effective communication and information delivery.

Moreover, the effectiveness of relation extraction is continually being enhanced through the integration of advanced machine learning techniques, such as deep learning. These techniques enable more accurate and nuanced recognition of relationships, even in complex sentences with multiple entities and relationships. The ongoing research and development in this field are directed towards making relation extraction models more robust, adaptable, and capable of handling a variety of linguistic expressions and structures.

In conclusion, Relation Extraction is not just a component of computational linguistics; it is a driving force behind many innovative applications in NLP [135]. Its ability to accurately interpret and categorize relationships between entities is fundamental to the advancement of language understanding technologies. The ongoing improvements in this area are continually expanding the boundaries of what machines can understand and achieve in the realm of natural language, marking it as a field of critical importance and immense potential within computational linguistics and NLP [136].

2.5.6 Data Augmentation using LLMs

From a data perspective, we group existing studies on LLM-based DA into four categories: 1. *Data creation* which leverages the few-shot learning ability of LLMs

to create a large synthetic dataset; 2. *Data labeling* which uses the LLM to label existing datasets; 3. *Data reformation* which transforms existing data to produce new data; 4. *Co-Annotate* which enables LLM-human collaboration to gather high-quality augmentation data. This section discusses relevant papers in each category.

2.5.6.1 Data Creation

Data Creation focuses on leveraging the few-shot learning ability of LLMs to quickly create a large amount of synthetic data. It is most used in tasks with a large label space. Data Creation with LLMs is a promising solution in specialized or private domains, where annotations are usually difficult or expensive to collect. Dialogue tasks are one example where specialized data is hard to collect. In medical dialogue summarization, Chintagunta et al. [137] uses a powerful few-shot learner such as GPT-3 to create synthetic medical dialogue summaries. By training models on a mix of synthesized and human-labeled data, the algorithm can scale a few human-labeled examples to yield results comparable to using 30x human-labeled examples. Similarly, for general dialogue, Dialogic [138] is seeded with a few dialogues and can automatically select in-context examples for demonstration and prompt LLMs to generate annotated dialogues in a controllable way. Then, automatic verification and revision methods are proposed to mitigate annotation errors. Results show that performance greatly improves in low-resource scenarios. Wan et al. [139] also attempt few-shot data augmentation on dialogue modeling. Aside from few-shot learning, for emotional support conversations, AugESC [140] finetunes an LM and prompts it to complete dialogues from collected posts. The post-training on AugESC improves downstream dialogue models' generalization abilities to open-domain topics. For low-resource classification, LLM can be used to create synthetic examples of a given label. Møller et al. [141] gives an example and its corresponding label and instruct the LLMs to generate similar examples exhibiting the same label. Resulting models yield better downstream performances on few-shot classification but still lag behind human-annotated data. For other low-resource tasks such as recommendation and intent detection, Data Creation can also effectively boost the training data space. To gather better recommendations, Zhang et al. [142] generates a large amount of user-personalized instruction data with varying preference and intention types. Then, the LLM is optimized using instruction tuning. The

resulting model can obtain more accurate recommendations and outperform competitive baselines, including GPT-3.5. For intent detection, Lin et al. [143] first uses an LLM to generate synthetic examples in the context of the training set and then uses Pointwise V-Information (PVI) to filter unhelpful examples. Sahu et al. [144] prompts GPT-3 to generate labeled training data, which can significantly boost the intent classifier's performance for distinct intents but becomes less helpful with semantically close intents.

Data Creation also helps in more general tasks by generating new training datasets. For information retrieval, Bonifacio et al. [145] uses few-shot prompting with LLMs to generate synthetic training datasets consisting of query-document pairs. Retrieval models finetuned with the augmented data significantly outperform unsupervised models. For reasoning, Logi-CoT [146] gathers a new instruction-tuning dataset by prompting GPT-4 and is used for teaching models to elicit general reasoning skills. Moreover, Data Creation is helpful for model performance distillation. To distill LLMs' reasoning performances to smaller models, Fine-tune-CoT [147] uses zero-shot CoT prompting to generate rationales from teacher models and use them to fine-tune smaller student models. The resulting performance improvements are stable across dataset size, teacher performance, etc. To reduce the need for manual annotation in reasoning tasks, Automate-CoT [148] automatically generates pseudo-CoTs from a small labeled dataset and then prunes and selects an optimal combination for CoT prompting. Similarly for instruction-following, Peng et al. [149] uses GPT-4 to generate an instruction-following dataset and feedback data. The resulting instruction-tuned LLaMa models can lead to comparable performance with the original GPT-4. To aid multilingual commonsense reasoning tasks, Whitehouse et al. [150] provides LLMs with instructions and examples from the original training data, prompting them to generate new and diverse examples. By training with augmented data, significant cross-lingual performance improvements are observed on smaller models.

To systematically study the behavior of such data creation methods and improve upon current few-shot prompting methods, Meng et al. [151] attempts to first tune an LM on few-shot examples and then use it as a generator to synthesize a large amount of novel training samples. The resulting approach could augment task performances than existing few-shot learning methods.

2.5.6.2 Data Labeling

Data Labeling seeks to utilize the general language comprehension abilities of LLMs to annotate unlabeled datasets. It is primarily useful in tasks that have a large enough unlabeled data corpus, such as cross-lingual and multimodal tasks. To evaluate LLMs' potential in data labeling, Törnberg [152] studies the zero-shot annotation ability of GPT-4 on labeling political Twitter messages with political tendencies. Compared to human workers, the LLM annotations display higher accuracy and lower bias. Similarly, Zhu et al. [153] observes that, in social computing tasks, ChatGPT has the potential to accurately reproduce human labels. Notably, annotations from open-source LLMs [154] and ChatGPT [155] can surpass crowd-worker performance on annotation tasks. For annotating low-resource tasks such as goal-oriented dialogues [156] and speech emotional data [157], the quality of ChatGPT annotations is on par with human-generated labels. However, Bansal and Sharma [158] observes that simply annotating can sometimes worsen generalization. Thus, it proposes conditional sampling to optimize the tradeoff between informativeness and budget.

Cross-lingual tasks mostly contain a large unlabeled corpus, which could benefit from data labeling. Therefore, Zhang et al. [159] uses different prompting strategies to augment machine translation (MT) data. It tries augmenting monolingual data using back-/forward-translation via zero-shot prompting, which still suffers from limitations such as generalization and unstable transfer performances. Similarly, Meoni et al. [160] annotates training data for multilingual clinical entity extraction with LLMs. After fine-tuning smaller models with augmentations, they display promising results for information extraction (IE) tasks.

Data labeling is also promising for multimodal applications. For data-scarce Visual Question Answering (VQA) tasks, Khan et al. [161] utilizes a Self-taught Data Augmentation (SelTDA) framework to generate pseudo labels from unlabeled images. The pseudo-labeled data could then improve VQA task performance and robustness. Combining multi-modality with reasoning, T-SciQ [162] further distills LLMs' reasoning abilities in multimodal tasks by asking the teacher model to produce CoT rationales. As a result, it achieves state-of-the-art performance in scientific QA.

2.5.6.3 Data Reformation

Data Reformation techniques attempt to reformulate the existing data into more variations for more fine-grained augmentation. Such reformation techniques could naturally aid in counterfactual generation tasks, which reforms existing data to its counterfactual version. Disco [163] uses LLMs to generate high-quality counterfactual data at scale. It first uses in-context learning with GPT-3 to generate phrasal perturbations, then uses a task-specific teacher model to filter and distill high-quality counterfactual data. Models trained using the generated counterfactuals display improved robustness and generalization across distributions. For retrieval-augmented generation, CORE [164] uses GPT-3 to generate counterfactual edits to the input conditioned on the retrieved excerpts. The perturbations then help mitigate model bias and improve performance on out-of-distribution (OOD) data.

Data Reformation could also quickly diversify the original dataset by forming data pairs. For conspiracy detection, Korenčić et al. [165] asks GPT-3 to rephrase tweets with original labels to augment training. To generate useful variations of the pre-training datasets of large vision models, ALIA [166] uses LLMs to generate image descriptions and augment the training data via language-guided image editing. By leveraging LLMs to the image domain, ALIA surpasses traditional data augmentation methods on fine-grained classification tasks. For Named Entity Recognition (NER), Sharma et al. [167] generates paraphrases while retaining inline annotation for entities. Among other PLMs, GPT-3 is

able to generate high-quality paraphrases, yielding statistically significant improvements in NER performance.

For more general tasks, Data Reformation could help to diversify and broaden the original dataset. AugGPT [168] tries to overcome the challenge of few-shot and data-scarce NLP tasks by rephrasing each sentence in the training samples into 6 semantically similar sentences. Experiments show that such an approach surpasses state-of-the-art text data augmentation methods in augmentation distribution and testing accuracy. For effective knowledge distillation, GPT3Mix [169] extracts sample sentences from the task-specific training data, embeds these samples in the prompt, and asks the LLM to generate an augmented mixed sentence influenced by the sample sentences. Guo et al. [170] asks GPT-3.5 and GPT-4 to rewrite or generate question-answer pairs with zero-shot prompting. Fine-tuning with

the refined and diversified training set then successfully distils medical question-answering abilities to smaller models.

2.5.6.4 Co-annotation

Co-annotation refers to the collaborative annotation process between humans and LLMs. By combining both annotation approaches, Co-annotation can reduce annotation costs and improve annotation performance at the same time. Firstly, Li et al. [171] proposes CoAnnotating, which allocates a given datapoint to be annotated by humans or by LLMs by computing the uncertainty level of LLM's annotations. With efficient human-AI collaboration, it provides insights into the tradeoff between annotation quality and annotation cost. To assist human annotators with explanations, Bertaglia et al. [172] asks the LLM to identify relevant features, such as text tokens, as assistive explanation. The approach improves inter-annotator agreement, annotation accuracy, and annotators' confidence, eventually leading to more transparency. Using human feedback to direct LLM annotations could also effectively generate high-quality data. Diagen [173] uses an LLM to iteratively generate dialogues in protected data domains, where human feedback is used to correct inconsistencies or redirect the flow in sub-dialogues. As a result, fine-tuning or in-context learning with the annotated data shows significant model performance improvements. Similarly, ToolCoder [174] uses human-written input-output pairs as prompts to guide chatGPT to annotate a tool-augmentation dataset. Then, the annotated data is filtered to ensure quality. After fine-tuning with the annotated data, ToolCoder can achieve comparable performance with ChatGPT on code generation.

Chapter 3

Generation-Based Data Augmentation Approach for Low-Resource Information Extraction

3.1 Background

A substantial volume of training data is crucial for the efficacy of neural models, particularly in the case of larger networks. Increasing the quantity of training data aids in mitigating overfitting and enhancing the robustness of the model. Nevertheless, the creation of a significant amount of annotated data typically involves considerable expense, extensive labor, and a lengthy process. In the realm of synthetic data generation, data augmentation [175] emerges as a valuable strategy. This technique is extensively applied in fields like computer vision [176–178] and speech recognition [179, 180].

While it's relatively straightforward to implement data augmentation methods in fields like computer vision and speech through techniques like rotation, cropping, and masking, these methods are more difficult to apply in natural language processing (NLP). This complexity arises because languages don't lend themselves to a set of general, handcrafted rules for transformation. In contrast to visual data, where simple distortions typically don't alter the underlying meaning, in language

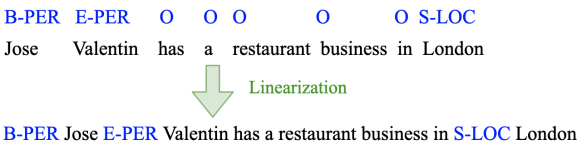


FIGURE 3.1: A demonstration of sentence linearization with labels involves coupling each word with its corresponding tag, typically by positioning the tags either before or after the words. In this process, the *O* tags are omitted.

processing, even a minor change such as deleting or replacing a single word can drastically shift the entire meaning of a sentence.

A notable strategy for data augmentation in Natural Language Processing (NLP) is 'back translation' [181–184]. This involves employing a translation model to convert sentences from the target to the source language, creating synthetic parallel sentences, namely back translation. Additional effective techniques comprise the systematic rearrangement of dependents in gold data to produce synthetic data for dependency parsing [185], utilizing knowledge bases for question generation [186], and adopting simulation-based methods to create a series of foundational toy tasks for QA [187]. Furthermore, methods like synonym substitution, random deletion, swapping, insertion, and generation using Variational Autoencoders (VAE) or pre-trained language models are frequently applied in various NLP tasks [18, 188–191], particularly in translation and classification.

In contrast to downstream tasks such as translation and classification, sequence tagging demonstrates increased sensitivity to data augmentation disturbances, owing to its detailed (token-level) nature. Employing methods like annotating unlabelled data with a basic tagger, utilizing parallel bilingual texts to generate annotations, and applying synonym substitution are three explored strategies for enhancing sequence tagging [192–194]. However, these approaches have drawbacks. The use of weakly labeled data typically results in additional noise. Notably, tagging unlabeled data with a rudimentary tagger demands both domain-specific data and expertise, to avoid the risk of domain-shift issues [195]. The application of parallel bilingual texts necessitates extra resources that might not be present for less common languages. Synonym replacement generally depends on supplementary information sources like WordNet [196], a manually curated lexicon that may lack comprehensive coverage for less widely spoken languages.

In this chapter, we explore the use of data augmentation via a generative approach for sequence tagging tasks. Initially, we convert the labeled sentences into a linear format, as illustrated in Figure 3.1. Following this, a language model (LM) is developed based on the linearized data, enabling the creation of synthetic labeled data. This approach differs from the use of weak taggers for labeling new data, as it integrates the processes of sentence creation and labeling through a LM. Specifically, we train the LM to generate a word and its associated tag simultaneously (for example, "B-PER Jose"), selecting tag-word pairs with a higher likelihood during generation. Our technique does not depend on external resources like WordNet. However, it remains adaptable to incorporate resources like unlabeled data or knowledge bases, employing a straightforward yet efficient conditional generation method.

While recent studies [18, 190, 191] have applied LM in data augmentation, their techniques are specific to sentence-level tagging, making them suitable only for classification tasks. In the domain of natural language processing (NLP), sentence-level augmentation strategies, while beneficial in various tasks such as text classification, are often inapplicable to sequence labeling tasks like named entity recognition (NER) and part-of-speech (POS) tagging. This inapplicability arises because sentence-level augmentations typically alter the structure or meaning of the entire sentence, potentially disrupting the sequential dependencies and contextual integrity crucial for accurate sequence labeling. For instance, an augmentation strategy that rephrases a sentence might change the position or context of entities, thereby confusing the labeling model. Consider the sentence "Barack Obama was born in Hawaii." A synonym replacement augmentation that changes "born" to "originated" can obscure the entity "Hawaii" as a place of birth, leading to incorrect labeling.

Contrary to these approaches, our work is pioneering in using generative language models to create detailed synthetic data from the ground up for sequence tagging tasks. This innovation presents a novel approach in data augmentation within the field of NLP. Additionally, our methodology isn't dependent on large, pre-trained models. Instead, we use a straightforward, single-layer recurrent language model [197], which is easier to train. The performance of our method is promising, showing significant results with training on a limited dataset of just a few thousand sentences.

In order to assess the efficacy of our approach, we carried out a wide range of experiments across various sequence tagging tasks such as named entity recognition (NER), part-of-speech (POS), and end-to-end target based sentiment analysis (E2E-TBSA). Our approach shows superior performance compared to the baseline methods in both supervised and semi-supervised scenarios. Unlike the baseline methods, our approach creates new synthetic data from the ground up, thereby adding more variety to minimize overfitting. In semi-supervised settings, our approach exhibits a robust capacity to utilize valuable insights from unlabeled data and knowledge bases.

3.2 Task Introduction

Named Entity Recognition (NER) Named entities are phrases in text that represent the names of people, organizations, locations, and similar entities. An example of this is, "*[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].*" The process of identifying and categorizing named entities in text into predefined categories is a crucial aspect of information extraction, commonly known as named entity recognition (NER) [198–200]. This task is notably challenging for a couple of reasons [100]: firstly, for NER, there is often a scarcity of manually annotated training data across various languages and domains; secondly, the limited training data makes it hard to achieve broad generalization due to restrictions on the types of words that can serve as names.

Part-of-Speech (POS) Tagging In the field of natural language processing (NLP), part-of-speech (POS) tagging plays a crucial role. This process involves assigning a grammatical category to each word in a sentence. Such tagging is essential for NLP systems, aiding in essential operations like syntactic parsing [201] and fine-grained sentiment analysis [202]. The latest POS tagging models have shown impressive accuracy, exceeding 97.80% on the PTB-WSJ dataset [104, 203] and averaging over 96.50% in test accuracy across 21 high-resourced languages in the UD 1.2 framework [204]. Nevertheless, these models face challenges, particularly a notable drop in accuracy when dealing with languages that have limited resources and uncommon words [105, 205].

Target Based Sentiment Analysis Target-based sentiment analysis is an essential component of sentiment analysis, focusing on identifying sentiment targets within sentences and determining the sentiment expressed towards these targets [202, 206–208]. Consider the statement, "USB3 Peripherals are noticeably less expensive than the ThunderBolt ones." Here, the sentence highlights two sentiment targets: "USB3 Peripherals" and "ThunderBolt ones," with the former receiving positive sentiment and the latter negative. Li et al. (2019) introduced an end-to-end approach for target-based sentiment analysis (E2E-TBSA), transforming this analysis into a tagging task. This method simultaneously addresses target identification and sentiment classification by assigning unified tags. For instance, the tag "B-POS" signals the start of a positively viewed target. Thus, the given example is annotated as, "[B-POS USB3] [E-POS Peripherals] are noticeably less expensive than the [B-NEG ThunderBolt] [E-NEG ones]" [208, 209].

3.3 Proposed Method

We introduce an innovative approach for enhancing data in sequence tagging assignments. Initially, we transform labeled sentences into a linear format, after which we employ a language model to comprehend the word and tag distribution within these linear sequences. This aids in creating artificial training material. Additionally, we suggest a method for conditional generation that leverages unlabelled data and knowledge bases when accessible.

3.3.1 Linearized Labeled Sentence

Initially, we conduct sentence linearization, transforming labeled sentences into sequential formats. This allows language models to effectively understand and learn the patterns of words and tags present in the original, high-quality data. As demonstrated in Figure 3.1, during this linearization process, tags are placed before their respective words, serving as their modifiers. In cases involving tasks where O tags are common, such as NER and E2E-TBSA [208], we exclude these tags from the linear sequences. In a similar vein, it's possible to position tags subsequent to their related words.

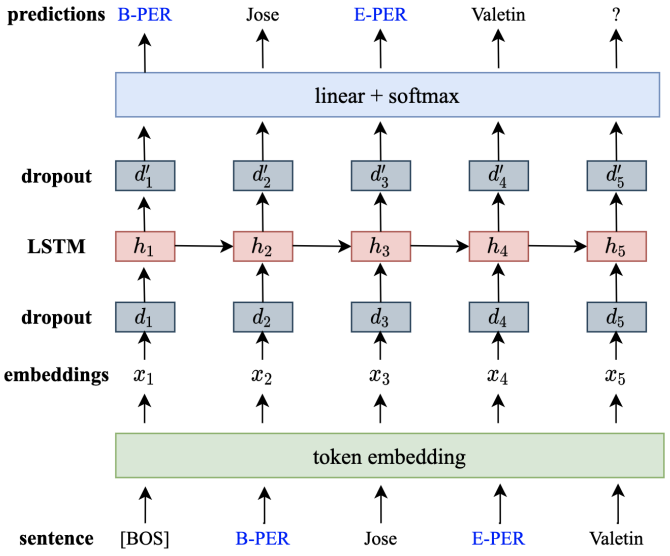


FIGURE 3.2: Language model architecture with LSTM.

Following the linearization of sentences, we introduce unique tokens, *[BOS]* at the start and *[EOS]* at the end of every sentence.

3.3.2 Data Generation via Language Modelling

When sentences with labels are linearized, language models become effective in understanding the patterns of word and tag distributions. In our approach, we employ a one-layer LSTM (Long Short-Term Memory) recurrent neural network language model (RNNLM), similar to the model introduced by [197]. The architecture of our RNNLM is depicted in Figure 3.2.

Language Modeling In the training of Recurrent Neural Network Language Models (RNNLM), the objective is to optimize the model for enhanced accuracy in predicting subsequent tokens. This process initiates with the input of a sentence, represented by a sequence of tokens $(w_1, w_2, \ldots, w_N)$. These tokens are then processed through an embedding layer, which maps them to their respective embeddings $(_{1,2}, \ldots, _N)$, with N denoting the length of the sequence. Subsequently, a dropout layer is employed on each token embedding $_t$, resulting in the generation of $_t = (_t)$. Following this, the sequence $(_{1,2}, \ldots, _N)$ is inputted into a Long Short-Term Memory (LSTM) network. This LSTM network is responsible for producing

[BOS] [labeled] B-PER Jose E-PER Valentin has a restaurant business in S-LOC London [EOS]

[BOS] [unlabeled] I have booked a flight to New York [EOS]

[BOS] [KB] We ate crepes in S-LOC Shibuya, saw cherry blossom bloom at Asakusa [EOS]

FIGURE 3.3: This illustration presents a case of conditional generation. The initial sequence originates from a dataset with verified named entity recognition (NER) annotations. The subsequent sequence is derived from a dataset devoid of labels, hence the absence of annotations. In the final sequence, labels are assigned based on matches with a knowledge base. However, the term 'Asakusa' remains unlabeled, indicative of gaps in the knowledge base's coverage.

a hidden state $_t = (_{t,t-1})$ at each sequential position t. To further enhance the model, another dropout layer is applied to these hidden states, thereby computing $'_t = (_t)$.

In the proposed framework, the final prediction of the subsequent token within the sequence is accomplished through the application of a linear layer combined with a softmax function. This process is contingent on the premise that the position of a given token w_t within the vocabulary is denoted as i^*. The training objective is succinctly encapsulated in Equation 3.3:

$$_{t-1} = ^{\mathsf{T}} '_{t-1} \tag{3.1}$$

$$p_\theta(w_t|w_{<t}) = \frac{\exp(_{t-1,i^*})}{\sum_{i=1}^{V} \exp(_{t-1,i})} \tag{3.2}$$

$$p(w_1, w_2, \ldots, w_N) = \prod_{t=1}^{N} p_\theta(w_t|w_{<t}) \tag{3.3}$$

Within this context, V represents the vocabulary's total size, $\in \mathbb{R}^{r\times V}$ symbolizes a trainable weight matrix, with r being the dimension of the LSTM hidden states. Moreover, $_{t-1,i}$ signifies the i-th component of $_{t-1}$.

Data Generation Upon completion of the Recurrent Neural Network Language Model (RNNLM) training, it becomes feasible to employ this model for the generation of synthetic training data applicable to tagging tasks. The generation process initiates with the exclusive input of the *[BOS]* token into the RNNLM. Subsequently, the sequence of tokens is determined through sampling, with each token's likelihood of selection deriving from the probabilities calculated as per

Equation 3.2. Commencing with *[BOS]*, sentence formation proceeds in an autoregressive manner, utilizing the token generated in the preceding step as the input for the subsequent token generation.

Equation 3.2 elucidates that during the generation phase, the RNNLM exhibits a propensity to select tokens that manifest higher probabilities. The inherent randomness of the sampling process enables the RNNLM to opt for tokens that are contextually similar. Consider a scenario where tags are inserted prior to their corresponding words in the process of sentence linearization for the purpose of RNNLM training. In this context, when the RNNLM is tasked with predicting the next token following the input *"I have booked a flight to"*, the probability of generating *"S-LOC"* is markedly higher compared to other options. This is attributed to the RNNLM's exposure to a multitude of analogous instances in the training data, such as *"a train to S-LOC"*, *"a trip to S-LOC"*, and so forth. Subsequent to this, the model engages in predicting the word that follows *"I have booked a flight to S-LOC"*. In the training dataset, every instance of *"S-LOC"* is succeeded by a location-specific word. Consequently, words like *"London"*, *"Paris"*, *"Tokyo"*, etc., emerge as viable selections, with their respective probabilities being comparably similar. The randomness incorporated into the model permits the selection of any of these options. In scenarios where tags are appended post the corresponding words, token prediction unfolds in an analogous manner, with the distinction that words are anticipated prior to the tags.

3.3.3 Post-Processing

The sequences produced are initially in a linearized configuration and necessitate reformatting to align with the structure of the benchmark data. Furthermore, this study proposes a set of basic yet effective rules for purifying the generated dataset: 1) Elimination of sentences devoid of tagging; 2) Removal of sentences consisting exclusively of the placeholder *[unk]*[1]; 3) Exclusion of sentences with erroneous sequencing of tag prefixes (for instance, the presence of *E-LOC* preceding *B-LOC* in Named Entity Recognition (NER) datasets); 4) Discarding sentences that exhibit identical word sequences but bear disparate tags.

[1]To streamline the vocabulary for language model training, infrequent words appearing only once in the training corpus are substituted with the *[unk]* token.

3.3.4 Conditional Generation

In this chapter, we introduce a novel method for conditional generation, designed to enhance language models through the utilization of unlabeled data or knowledge bases, particularly in scenarios where resources are limited. For instance, annotating substantial volumes of e-commerce product titles for Named Entity Recognition (NER) can be prohibitively costly, whereas acquiring a knowledge base (such as a dictionary) of product attributes and unlabeled data is more feasible. Our approach involves prefixing each sequence with one of the designated **condition tags** {*[labeled]*, *[unlabeled]*, *[KB]*} to identify its source. Here, *KB* indicates that the sequence has been labeled through a comparison of a knowledge base with unlabeled data, as illustrated in Figure 3.3. This strategy enables the language model to assimilate shared information across these sequences while recognizing their distinct origins. In the process of generating synthetic data, each word is generated in relation to the specified condition tag *[labeled]*, symbolized as c (conditioning class). This tag is then incorporated into the language model, resulting in all subsequent Long Short-Term Memory (LSTM) hidden states containing information about c. Moreover, also encapsulates details from all preceding tokens in the sequence. Consequently, when predicting the succeeding token based on $_{t-1}$, the probability $p_\theta(w_t|w_{<t})$ in Eq. 3.2 is modified to $p_\theta(w_t|w_{<t}, c)$.[2] This methodology aligns with the approach employed in CTRL [210], which is utilized to modulate style and task-specific behaviors in text generation.

3.4 Experiments

In this section, the experimentation conducted under supervised and semi-supervised frameworks is delineated. Within the confines of the supervised framework, the augmentation process exclusively utilizes gold data (ground-truth data). Conversely, the semi-supervised framework extends its reliance to incorporate both unlabeled data and knowledge bases, in addition to the gold data.

[2]Notably, the condition tag c is also part of $w_{<t}$ as it is a unique token added at the start of each sentence. However, it is explicitly mentioned here to highlight its conditional influence.

3.4.1 Base Models

Language Model In this chapter, the synthetic data generation process employs the language model delineated in Section 3.3.2. Adaptations were made to the decoder of the LSTM-LM model, as detailed in [211], to facilitate the implementation of this language model. The LSTM hidden state and embedding sizes were configured to 512 and 300, respectively. For regularization, a dropout rate of 0.5 was applied to both dropout layers. The training of all language models utilized Stochastic Gradient Descent (SGD) with an initial learning rate of 1 and a batch size of 32. The learning rate was subject to a decay factor of 0.5 in subsequent epochs, contingent upon the absence of improvement in perplexity on the development set. The training process was capped at a maximum of 30 epochs, with an early termination clause activated if there was no improvement in perplexity on the development set across three consecutive epochs. In the synthetic data generation phase, the average length of gold sentences from the training set was used as the benchmark for maximum sentence length.

Sequence Tagging Model In the experiments, a BiLSTM-CRF model, as described by [100], was deployed within the context of the Flair framework [212] to assess the effectiveness of a novel data augmentation technique on Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks.[3] The implemented model consists of a single-layer BiLSTM, equipped with a hidden state size of 512. To mitigate overfitting, dropout layers were integrated both preceding and succeeding the BiLSTM layer, maintaining a dropout rate of 0.5. Training of all sequence tagging models was conducted using the Adam optimizer [213], starting with an initial learning rate of 1e-3 and a batch size of 32. The learning rate was subject to a 0.5 reduction if no performance improvement was observed on the development set over three consecutive epochs. Training was terminated either when the learning rate fell below 1e-5 or upon reaching 100 epochs. For all languages, the pre-trained 300-dimensional fastText word embeddings [214] were employed.

We utilize fundamentally simple basic models for several reasons. Firstly, these models aid in mitigating the risk of overfitting, a concern that arises from the limited data availability inherent in low-resource settings. Secondly, they facilitate

[3]For evaluation of the end-to-end target-based sentiment analysis task, the baseline model from the original study was utilized.

a more accurate comprehension of the impacts attributed to the data augmentation method we propose.

3.4.2 Supervised Experiments

To assess the efficacy of our proposed data augmentation technique within supervised learning contexts, we conducted evaluations across three distinct tagging tasks: Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and End-to-End Task-Based Sentiment Analysis (E2E-TBSA). In contrast to many existing studies that incorporate supplementary data, our approach was benchmarked against a baseline method of random deletion (**rd**), as described by Wei and Zou [189]. This method entails the arbitrary elimination of 5% of words[4] along with their associated tags from the training dataset. For a comprehensive overview of the methodologies employed in our supervised experiments, refer to Table 3.1.

Method	Description
gold	This approach exclusively utilizes the gold data.
gen	This is our proposed methodology. It involves the generation of synthetic data utilizing language models, coupled with an oversampling of the gold data.
rd	A baseline approach. It generates synthetic data through a process of random deletion, followed by oversampling of the gold data in a ratio identical to that used in the **gen** method.
rd*	Another baseline approach, bearing resemblance to **rd**. However, it differentiates itself by ensuring equal sampling proportions between the gold and synthetic data.

TABLE 3.1: Data sources for the supervised setting.

3.4.2.1 Named Entity Recognition

Dataset The efficacy of our proposed methodologies was evaluated on the CoNLL2002/2003 NER datasets [215, 216], encompassing four distinct languages: English, German, Dutch, and Spanish. Additionally, the evaluation extended to

[4]In the cases of NER and E2E-TBSA, if a selected word is part of an entity span, the entire entity is removed.

NER datasets in Thai and Vietnamese, specifically focusing on product titles sourced from leading e-commerce platforms in Southeast Asian nations. These datasets were meticulously annotated with eleven product attribute NER tags, namely *PRODUCT*, *BRAND*, *CONSUMER_GROUP*, *MATERIAL*, *PATTERN*, *COLOR*, *FABRIC*, *OCCASION*, *ORIGIN*, *SEASON*, and *STYLE*. The comprehensive statistics pertaining to the Thai and Vietnamese NER data utilized in our experiments are delineated in the Appendix.

Experimental Settings To assess our methodology, we not only used the complete training dataset but also took random samples of 1k, 2k, 4k, 6k, and 8k sentences from each language to check its effectiveness in situations with limited resources. The training involved using these sampled sentences to create synthetic training data, as outlined in Section 3.3, while retaining the original development and test data.

We then produced 1k-sentence batches using the trained language models. In these batches, we tracked the emergence of new 1-gram tokens. Data generation ceased when 99% of these tokens were found in prior batches. Following this, we refined the data as per Section 3.3.3 and integrated it into the primary training set for the tagging model. For the **rd** and **gen** scenarios, we enhanced the gold data by fourfold repetitions in a shuffled order within the training set. In cases of random deletion, we also evaluated scenarios where gold and synthetic data were utilized in equal measure, indicated as **rd***. Further information about the parameters used for determining these oversampling ratios is available in Appendix A.3. Consistent with [100], we employed the IOBES tagging system for training both the language and sequence tagging models.

Results and Analysis The summarized findings are presented in Table 3.2, which reflects the mean outcomes from three trials. Across all languages, our approach consistently enhanced performance. This improvement is notably more pronounced in smaller sampled datasets. Specifically, in comparison to the baseline methods, our proposed technique attained an average enhancement of 1.93 and 1.38 points in the 1k and 2k configurations, respectively.

Lang.	Method	1k	2k	4k	6k	8k	all
en	gold	58.06	67.85	74.55	77.16	80.30	83.04
	+rd*	59.42	67.23	74.51	77.39	80.31	83.39
	+rd	58.97	67.81	74.77	77.35	80.59	83.25
	+gen	**61.15**	**70.61**	**76.82**	**79.18**	**81.02**	**83.74**
de[5]	gold	29.71	**41.07**	49.55	53.30	56.17	61.10
	+rd*	29.89	40.29	49.27	52.33	55.70	60.69
	+rd	30.83	40.36	49.24	53.54	55.60	60.55
	+gen	**31.83**	40.92	**49.79**	**53.63**	**56.94**	**62.44**
es	gold	58.14	67.42	74.21	77.44	78.90	79.27
	+rd*	58.22	66.98	75.08	77.64	79.11	80.01
	+rd	59.67	68.53	75.21	77.79	79.12	80.26
	+gen	**61.76**	**68.62**	**76.15**	**78.20**	**79.83**	**80.73**
nl	gold	37.04	48.61	57.78	61.08	64.59	70.89
	+rd*	35.10	46.45	56.83	60.49	63.09	69.42
	+rd	**39.39**	48.44	59.38	61.48	64.44	70.36
	+gen	38.87	**50.41**	**59.90**	**63.19**	**65.82**	**72.71**
vi	gold	55.98	62.42	69.01	70.75	72.12	76.14
	+rd*	55.67	63.57	68.47	70.87	72.08	76.43
	+rd	56.24	63.08	68.63	71.15	72.22	76.83
	+gen	**60.01**	**65.43**	**70.36**	**72.55**	**74.11**	**77.39**
th	gold	49.88	55.79	61.75	63.10	64.94	67.71
	+rd*	50.46	56.98	62.12	64.19	66.47	67.81
	+rd	50.52	57.42	61.51	64.59	66.07	67.97
	+gen	**54.02**	**59.36**	**63.94**	**66.21**	**68.05**	**69.86**

TABLE 3.2: Named entity recognition micro F1.

Tag-Word vs. Word-Tag Section 3.3.1 outlines two approaches for sentence linearization: firstly, placing tags before the words they correspond to (Tag-Word), and secondly, positioning tags after these words (Word-Tag). When other variables remain constant, Tag-Word demonstrates superior performance over Word-Tag in NER tasks, a finding detailed in Appendix A.2. This could be attributed to the alignment of Tag-Word with the commonly seen Modifier-Noun structure in language modeling training data. Hence, Tag-Word is the chosen method for all NER experiments.

3.4.2.2 Part of Speech Tagging

Dataset Our evaluation of this task utilizes Part-of-Speech (POS) data obtained from the Universal Dependencies treebanks[6]. This evaluation encompasses five languages: English, Spanish, Czech, Romanian, and Japanese. Within the Universal Dependencies treebanks, each of these languages is represented by several corpora. To create a consolidated dataset for three of the languages — English, Spanish,

[6]https://universaldependencies.org/

Lang.	Method	1k	2k	4k	6k	8k	Full
en	gold	79.18	82.17	85.83	88.62	90.21	93.00
	+rd*	79.28	82.42	85.82	88.55	90.07	92.89
	+rd	79.38	82.50	86.08	88.80	90.15	92.96
	+gen	**79.76**	**82.90**	**86.31**	**88.99**	**90.56**	**93.29**
es	gold	88.28	90.79	92.82	93.80	94.43	96.40
	+rd*	88.25	90.94	92.84	93.76	94.48	96.41
	+rd	88.17	90.78	92.79	93.67	94.28	**96.45**
	+gen	**88.77**	**91.04**	**93.12**	**93.93**	**94.64**	**96.45**
cz	gold	80.10	84.46	88.88	90.67	92.03	97.52
	+rd*	79.83	84.29	88.64	90.43	91.95	97.57
	+rd	80.11	84.50	88.99	90.66	91.86	97.60
	+gen	**80.65**	**85.17**	**89.58**	**91.22**	**92.49**	**97.63**
ro[7]	gold	86.69	89.57	92.73	93.84	94.54	94.54
	+rd*	86.42	89.58	92.50	93.89	94.64	94.64
	+rd	86.62	89.46	92.55	93.84	94.73	94.73
	+gen	**87.29**	**90.66**	**93.44**	**94.61**	**95.17**	**95.17**
ja[8]	gold	90.19	91.44	93.59	94.41	-	95.08
	+rd*	90.00	91.41	93.66	94.62	-	94.93
	+rd	89.53	91.76	93.62	94.59	-	95.18
	+gen	**91.00**	**92.51**	**94.12**	**95.21**	-	**95.45**

TABLE 3.3: POS tagging accuracy.

and Czech — we combine these multiple corpora. Specifically, for English, the corpora *GUM*, *ParTUT*, *PUD*, and *Lines* are merged. In the case of Spanish, the *AnCora* and *GSD* corpora are amalgamated. For Czech, the merger includes *PDT*, *FicTree*, *CLTT*, and *CAC*. Additionally, our model has undergone evaluation using Japanese (*GSD*) and Romanian (*RRT*) datasets, which represent languages either spoken by smaller populations or belonging to distinct language families.

Settings and Results Our approach adopts comparable experimental configurations as used in the NER task. We employ the identical language model and the BiLSTM-CRF model for creating synthetic data and conducting POS tagging, respectively. Unlike NER, the Word-Tag model exhibits a marginally superior efficacy in POS tasks (see Appendix A.2).

The mean Word-Tag outcomes across three trials are depicted in Table 3.3. Our strategy shows a uniform enhancement in performance across all languages. Echoing the patterns observed in NER, our approach marks a more pronounced improvement in performance, particularly for smaller subsets in POS tagging. Specifically, the introduced method registers an average increase of 0.56, 0.60, and 0.46 points over the baseline methods in the 1k, 2k, and 4k sample sizes, respectively.

3.4.2.3 Target Based Sentiment Analysis

Dataset The evaluation of E2E-TBSA utilizes laptop and restaurant review datasets, originally sourced from SemEval ABSA challenges as indicated by [116, 117, 217] and subsequently processed by [208]. We combine these datasets and designate 10% of the training data, selected at random, as the development set. For scenarios with limited resources, we create smaller subsets from the remaining training data. The test sets from the original data are consolidated to form our unified test set.

Settings and Results Our methodology mirrors the experimental framework used in NER and POS tasks, with the notable exception of employing the sequence tagging model provided by [208] for our evaluations. In this context, Tag-Word exhibits superior performance, as detailed in Appendix A.2. This enhanced performance is likely attributed to the uniformity of tags (such as *B-POS*, and *B-NEG*) that resemble noun modifiers, aligning well with the Modifier-Noun structure inherent in Tag-Word. The mean outcomes of Tag-Word across three iterations are displayed in Table 3.4. Our approach shows an uptick in effectiveness, particularly for datasets larger than 4k. It's noteworthy that, in contrast to the NER and POS datasets, the E2E-TBSA dataset comprises significantly fewer labels, leading to more variable results.

Method	2k	4k	all(6k)
gold	56.31	60.43	63.18
+rd*	**57.92**	61.75	63.66
+gen	57.07	**62.66**	**65.86**

TABLE 3.4: E2E-TBSA micro F1.

3.4.3 Semi-supervised Experiments

In this part, we assess how well our approach works under two semi-supervised scenarios: a) when only non-labeled data is accessible; b) when access to both non-labeled data and a knowledge base is present. Consult Table 3.5 for the symbols representing the techniques applied in our semi-supervised tests.

Method	Description
gold	Supervised method using only the gold data.
wt	Baseline method where a weak tagger (a model trained on the gold data) labels the unlabeled data.
gen$_{ud}$	Our method which creates synthetic data using a Language Model (LM) that is trained on both gold and unlabeled data.
kb	Baseline method where the unlabeled data is annotated using a knowledge base.
gen$_{kb}$	Our method where synthetic data is generated using a LM, which in this case is trained on both gold standard data and data annotated with a knowledge base.

TABLE 3.5: Data sources for the semi-supervised setting.

3.4.3.1 Only Using Unlabeled Data

Dataset In this part, we assess how well our approach works in two different semi-supervised contexts. Our evaluation relies on the CoNLL2003 English NER dataset [216]. Beyond the primary NER training data, we also incorporate unlabeled data for semi-supervised learning. For sentence tokenization, we employ the Stanford CoreNLP tokenizer [218], specifically applied to sentences from Wikipedia.

Experimental Settings In experiments comparable to those previously mentioned, we utilize varying quantities of sentences (1k, 2k, 4k, 6k, 8k, and the entire dataset) randomly selected from NER gold data for the evaluation of our approach. To ensure a level playing field, both our method and the control group use an identical set of 10k sentences, randomly chosen from a Wikipedia dump. Here, we refer to the NER gold data samples as D_{gold} and the Wikipedia samples as $D_{unlabeled}$.

In our approach, we combine D_{gold} and $D_{unlabeled}$ to train language models as outlined in Section 3.3.4. These models are then employed to create synthetic data, of which 20k sentences are selected at random and combined with D_{gold} for the training of NER models. This process of generating data is signified by **gen$_{ud}$**. As a comparison, we use a baseline method that involves annotating $D_{unlabeled}$ with weak taggers, referred to as **wt**. In this case, the weak taggers are NER models trained using D_{gold}. Both our method and the baseline utilize the same NER model (BiLSTM-CRF) and hyperparameters for evaluation purposes. During the language model training phase, sentences from D_{gold} and $D_{unlabeled}$ are sampled

Method	**1k**	**2k**	**4k**	**6k**	**8k**	**all**
gold	58.06	67.85	74.55	77.16	80.30	83.04
lightyellow +wt	65.12	72.43	77.90	79.41	81.36	84.00
lightyellow +gen_{ud}	**66.19**	**73.00**	**78.08**	**79.75**	**81.98**	**84.33**
lavender +kb	**67.36**	72.86	77.15	79.33	81.91	83.69
lavender +gen_{kb}	66.67	**73.54**	**78.32**	**79.98**	**81.93**	**84.03**

TABLE 3.6: Semi-supervised NER F1.

equally. For training the NER models, we amplify the presence of gold data by replicating D_{gold} four times, resulting in a mixed training file.

Results and Analysis F1 scores of **wt** and $\textbf{gen}_{\textbf{ud}}$ (calculated as the average of three trials) are presented in Table 3.6. The performance of our approach surpasses that of the standard **wt** method across various configurations. Furthermore, it appears to be a promising avenue to enhance our model's performance by integrating an increased volume of unlabeled data. The conventional **wt** method faces challenges in incorporating a substantial volume of unlabeled data, as this leads to a proportional increase in the augmented data, potentially leaving some of it unused before the sequence tagging models reach convergence. In contrast, our $\textbf{gen}_{\textbf{ud}}$ method can efficiently handle large quantities of unlabeled data, which significantly boosts the quality of data augmentation. With an abundance of unlabeled data available, our strategy involves initially pretraining the language models using this data and subsequently refining them with labeled data.

3.4.3.2 Using Unlabeled Data and Knowledge Base

Dataset In this experiment, we explore the use of a knowledge base to enhance performance, in addition to the gold training data and unlabeled sentences mentioned in Section 3.4.3.1. The knowledge base is constructed by identifying entities (which are case sensitive and appear a minimum of two times) and their respective tags from the entire gold training dataset. Furthermore, we enrich the knowledge base with additional *LOC* entities by incorporating cities and countries derived from the geonames database[9].

[9]https://datahub.io/core/world-cities

Experimental Settings In the approach outlined in Section 3.4.3.1, we employ a random selection process for both the gold NER dataset and Wikipedia content, designating these subsets as D_{gold} and $D_{unlabeled}$, respectively. The process for annotating $D_{unlabeled}$ involves utilizing our knowledge base to identify the longest forward matches within each sentence, a technique we refer to as **kb**, resulting in a dataset named D_{kb}.

In our methodology, we merge D_{gold} and D_{kb} to facilitate the training of language models, adhering to the protocol detailed in Section 3.3.4. These language models are then deployed to create synthetic data. From this synthetic data, a subset of 20,000 randomly chosen sentences is amalgamated with D_{gold} for the training of NER models. This approach of data generation, labeled as $\mathbf{gen_{kb}}$, is assessed in comparison to the benchmark method **kb**. Consistent with previous experiments, we implement an oversampling strategy for D_{gold} in both the language and NER model training processes.

Results and Analysis In Table 3.6, we display the F1 scores for **kb** and $\mathbf{gen_{kb}}$ (based on an average from three attempts). The **kb** baseline demonstrates exceptionally strong results when D_{gold} is limited in size. This is attributed to the extensive use of a comprehensive database of countries and cities for labeling, coupled with the fact that locations are generally less prone to ambiguity than other entity types. Nonetheless, when the volume of D_{gold} exceeds 2,000, our approach surpasses **kb**. This indicates our method's enhanced resilience to inaccuracies present in D_{kb}, particularly when a marginally larger dataset of gold data is available.

3.5 A Closer Look at Synthetic Data

In this part of the discussion, we delve deeper into the reasons why our approach's synthetic data is beneficial for enhancing sequence tagging results. By examining the generated data more closely, we've come across a number of noteworthy insights.

More Diversity The creation of synthetic data enhances diversity, which mitigates the risk of overfitting. Illustrated in Figure 3.4, in the original training data, the name "Sandrine" is consistently associated with "Testud" across various

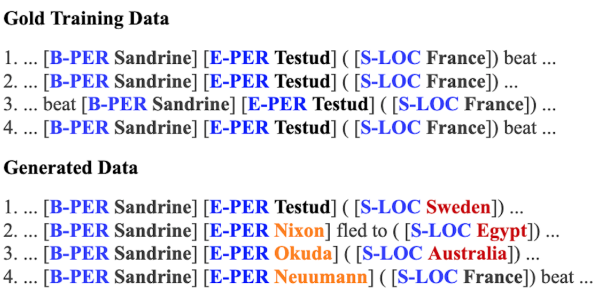


FIGURE 3.4: An illustration of diversity of generated data. The name "*Sandrine*" in the gold training data always pairs up with "*Testud*" in sentences.

sentences. Conversely, the synthetic data exhibits novel pairings such as "Sandrine Nixon," "Sandrine Okuda," and "Sandrine Neuumann." Additionally, the geographical references in the sentences are substituted with countries like "Sweden," "Egypt," and "Australia." This approach enables the model to concentrate on discerning the contextual patterns of entity occurrence rather than merely associating "Sandrine Testud" with a personal name and "France" with a location.

To objectively assess the diversity contributed by our technique and its effect, we performed a statistical examination of the contextualized entities (CEs) in the synthetic data for supervised English NER. A CE represents an entity and its immediate one-word context. For instance, in "The [B-ORG European] [E-ORG Commission] said ...", the entity "European Commission" and its CE is "The European Commission said". As depicted in Figure 3.5, we calculated the total unique CEs in the original training data and the number of new unique CEs in the synthetic data, along with their ratio. We also graph the improvement in F1 score using our method (gold+gen) compared to solely the original data, as indicated in Table 3.2. Our findings reveal that the method introduces a substantial number of new CEs, thereby enhancing the resilience of the resulting model. A higher ratio correlates with greater F1 improvement, demonstrating that our method effectively addresses the issue of scarce resources by generating a wealth of new entities and contexts. For additional details on unique entities (excluding context), refer to Appendix A.5, which corroborates our findings.

Efficient Usage of Unlabeled Data Our approach is adaptable in incorporating unlabeled data for semi-supervised training when such data is available. We've observed numerous intriguing instances in the synthetic data, demonstrating how our method efficiently leverages unlabeled data to glean valuable insights. Take

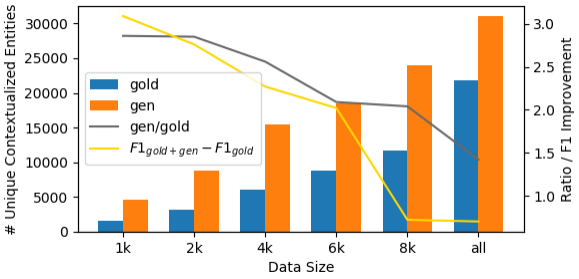


FIGURE 3.5: Statistics of unique contextualized entities.

the instance generated by our method, "... the [B-ORG Bank] [I-ORG of] [E-ORG Alabama] ...", where the term "Alabama" is not found in the original NER training dataset. Yet, our language model recognized "Alabama" (sourced from unlabeled data) as akin to other geographical names present in both the primary training dataset and unlabeled data. Therefore, when creating synthetic data, the language model can apply this term in comparable contexts or even invent new entities (for instance, "Bank of Alabama" is a novel entity not previously seen in any training data).

3.6 Chapter Summary

In the current chapter, we embark on a detailed exploration of the utility and efficiency of language models in the generation of high-quality synthetic data, specifically tailored for sequence tagging tasks in monolingual language contexts. The pivotal aspect of this approach lies in the creation of an entirely new dataset, which inherently introduces a diverse range of linguistic elements. This diversity plays a crucial role in addressing and substantially reducing the prevalent issue of overfitting, a challenge often encountered in machine learning models. Unlike traditional methods that primarily focus on modifying existing gold standard training materials, our strategy emphasizes the construction of this dataset from the very foundations.

This innovative method, which we have meticulously developed and proposed, has demonstrated significant improvements in the realm of sequence tagging tasks. The enhancements in performance are particularly noteworthy in situations where

resources are scarce or limited. This aspect is of paramount importance, as it highlights the adaptability and resourcefulness of our proposed method in overcoming the constraints of limited data availability.

Furthermore, our approach showcases a remarkable ability to integrate both unlabeled data and extensive knowledge bases effectively. This is achieved within a semi-supervised training framework, which is a testament to the flexibility and efficiency of our method. The incorporation of these elements is instrumental in enriching the training process, thereby leading to more robust and well-rounded machine learning models. This strategy not only broadens the scope of applicability of our approach but also enhances its effectiveness in diverse real-world scenarios.

In summary, this chapter provides a comprehensive examination of a novel approach to generating synthetic data for sequence tagging in monolingual languages. Through this method, we address key challenges such as overfitting and limited resource availability, while also expanding the potential of incorporating varied data sources in a semi-supervised learning environment. The encouraging results obtained from various tagging exercises underscore the efficacy and potential of our proposed methodology in advancing the field of language model applications.

3.7 Supplementary Materials

3.7.1 Statistics of Thai and Vietnamese NER Data

We present the number of sentences in Thai and Vietnamese NER data in Table 3.7.

Lang.	train	dev	test
vi	18,922	500	500
th	11,272	499	490

TABLE 3.7: Number of sentences in TH and VI NER data.

3.7.2 Experiments on Tag-Word vs. Word-Tag

We conduct experiments to compare the performance of Tag-Word and Word-Tag for the tagging tasks. All of the other settings are same as the corresponding

experiments presented in the main paper. Results are reported in Table 3.8 to 3.10. Tag-Word yields better average performance for NER and E2E-TBSA, while Word-Tag slightly outperforms Tag-Word for POS tagging.

Lang.	Method	1k	2k	4k	6k	8k	full	average
en	Tag-Word	**59.39**	**69.48**	**75.68**	**78.65**	**80.19**	**83.70**	**74.52**
	Word-Tag	58.97	67.32	75.45	78.06	80.43	83.58	73.97

TABLE 3.8: CoNLL NER F1: Tag-Word vs. Word-Tag.

Lang.	Method	1k	2k	4k	8k	15k	average
en	Tag-Word	79.06	82.43	85.93	**90.38**	**92.75**	86.11
	Word-Tag	**79.18**	**82.64**	**86.13**	90.33	92.68	**86.19**

TABLE 3.9: Universal Dependencies POS accuracy: Tag-Word vs. Word Tag.

Lang.	Method	2k	4k	full(6k)	average
en	Tag-Word	54.22	**61.72**	**62.88**	**59.61**
	Word-Tag	**55.58**	59.42	61.65	58.88

TABLE 3.10: E2E-TBSA micro F1: Tag-Word vs. Word-Tag.

3.7.3 Experiments on Oversampling Ratios

We conduct experiments to compare different oversampling ratios for NER task. Results are reported in Table 3.11. The notation gold$\times N$ means we oversample gold by repeating it N times in the shuffled static training data.

3.7.4 Semi-supervised Experiments on Part of Speech Tagging

Dataset We use the English POS data from Universal Dependencies treebanks for evaluation on this task. We merge *GUM*, *ParTUT*, *PUD* and *Lines* corpora to build the English dataset. Similar to the semi-supervised experiments on NER, we also utilize unlabeled Wikipedia sentences for training.

Experimental Settings We use 1k, 2k, 4k, 6k and 8k sentences randomly sampled from English POS gold data as well as the full dataset to evaluate our

Lang.	Method	1k	2k	4k	average
en	gold×1	59.74	69.14	76.48	68.45
	gold×2	60.92	69.79	76.57	69.09
	gold×3	61.13	70.42	74.92	67.24
	gold×4	61.15	**70.61**	**76.82**	**69.53**
	gold×5	**61.43**	70.38	76.43	69.41

TABLE 3.11: CoNLL NER F1: comparison on different oversampling ratios.

method. We follow the same experimental setting as the semi-supervised experiments on NER to generate synthetic data, train the sequence tagging models and evaluate on the POS test data.

Results and Analysis We report accuracy of **wt** and $\mathbf{gen_{ud}}$ (average of 3 runs) in Table 3.12. Our method outperforms the baseline method **wt** when the number of gold sentences are less than 8k. When the number of gold sentences are more than 8k, the performance of our method is comparable with **wt**.

Method	1k	2k	4k	6k	8k	all
gold	79.18	82.17	85.83	88.62	90.21	93.00
+wt	81.11	84.00	86.91	89.64	**90.88**	**93.20**
$+gen_{ud}$	**82.11**	**84.93**	**87.52**	**89.98**	90.84	93.12

TABLE 3.12: Semi-supervised POS accuracy.

3.7.5 Synthetic Data Diversity: Unique Entities

To quantitatively measure the diversity introduced by our method in the supervised English NER tasks, we count the number of unique entities (without context) in the gold and generated data. Results are presented in Figure 3.6.

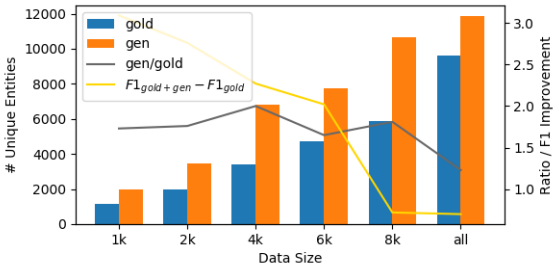


FIGURE 3.6: Statistics of unique entities (without context)

3.7.6 Average Runtime

Table 3.13 is an illustration of the average runtime of our models in English NER, POS and E2E-TBSA tasks and RNNLM.

Task	1k	2k	4k	6k	8k	all
NER	26.5	70.5	124.4	167.9	216.3	393.2
POS	83.6	112.3	231.1	257.7	277.2	298.0
E2E-TBSA	-	89.3	150.8	269.2	-	-
RNNLM	0.7	1.1	2.0	2.7	3.3	3.9

TABLE 3.13: Average runtime (min).

3.7.7 Computing Infrastructure

We conduct our experiments on NVIDIA V100 GPU.

Chapter 4

Developing a Task-Specific Multilingual Dialogue System Incorporating Local Entities and Contexts in a Global Setting

4.1 Background

In the preceding chapters, the discourse predominantly centered on the domain of information extraction, underscoring its role as a cornerstone in the array of fundamental Natural Language Processing (NLP) tasks. This exploration has laid a critical foundation for understanding the intricate mechanisms by which data is systematically retrieved and processed from natural language sources. Nevertheless, to fully appreciate the scope and potential of NLP, it is imperative to broaden our perspective beyond these foundational tasks and delve into the realm of downstream applications, most notably, dialogue systems. Such systems epitomize the practical implementation of NLP techniques, transforming theoretical concepts into tangible, user-interactive platforms. They exemplify the evolution of NLP from mere information extraction to the creation of dynamic, responsive, and intelligent communication interfaces, thereby representing a significant leap in the practical application of NLP methodologies.

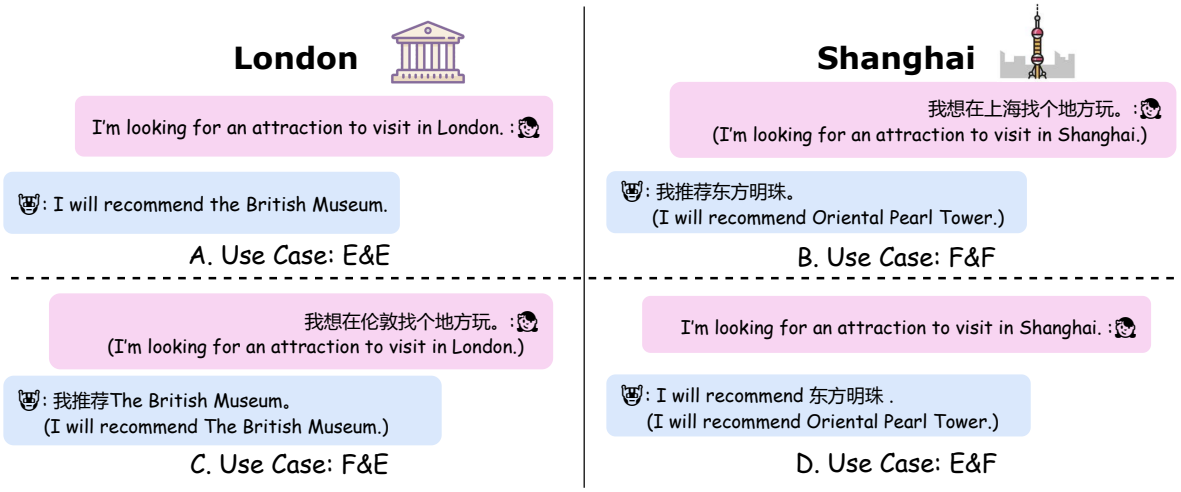


FIGURE 4.1: Examples of four use cases for multilingual ToD systems: A. Use Case E&E: A English speaker travels to a country of English. B. Use Case F&F: A foreign language speaker travels to a country of the foreign language. C. Use Case F&E: A foreign language speaker travels to a country of English. D. Use Case E&F: A English speaker travels to a country of a foreign language.

One of the fundamental objectives in pursuit of artificial intelligence is to enable machines with the ability to intelligently communicate with human in natural languages, with one of the widely-heralded applications being the task-oriented dialogue (ToD) systems [219, 220]. Recently, ToD systems have been successfully deployed to assist users with accomplishing certain domain-specific tasks such as hotel booking, alarm setting or weather query [221–224], thanks to the joint advent of neural networks and availability of domain-specific data. However, most existing ToD systems are predominately built for English, limiting their service for *all* of the world's citizens. The reason of this limitation lies in the stark lack of high-quality multilingual ToD datasets due to the high expense and challenges of human annotation [21].

Moreover, the challenge of localization in NLP with respect to cultural context is exemplified by the difficulty in adapting sentiment analysis tools across different cultural backgrounds. Sentiments expressed in one culture may carry different connotations in another, making it challenging to maintain accuracy without cultural adaptation. For instance, the phrase "break a leg," commonly used in Western cultures to wish someone good luck, might be misinterpreted as a literal negative expression in other cultures. Thus, without localization that respects cultural nuances, NLP models can fail to accurately interpret and process language in a culturally sensitive manner.

One solution to this is annotating conversations in other languages from scratch, e.g., CrossWoZ [225] and BiToD [226]. However, these methods involve expensive human efforts for dialogue collection in the other languages, resulting in a limited language/domain coverage. The other major line of work focused on translating an existing English ToD dataset into target languages by professional human translators [227–230]. Despite the increasing language coverage, these methods simply translated English named entities (e.g., location, restaurant name) into the target languages, while ignored the fact that these entities barely exist in countries speaking these languages. This hinders a trained ToD system from supporting the real use cases where a user looks for local entities in a target-language country. For example in Figure 4.1, a user may look for the British Museum when traveling to London (A.), while look for the Oriental Pearl Tower when traveling to Shanghai (B.).

In addition, prior studies [231, 232] have shown that code-switching phenomena frequently occurs in a dialogue when a speaker cannot express an entity immediately and has to alternate between two languages to convey information more accurately. Such phenomena could be ubiquitous during the cross-lingual and cross-country task-oriented conversations. One of the reasons for code-switching is that there are no exact translations for many local entities in the other languages. Even though we have the translations, they are rarely used by local people. For example in Figure 4.1 (C.), after obtaining the recommendation from a ToD system, a Chinese speaker traveling to London would rather use the English entity "British Museum" than its Chinese translation to search online or ask local people. To verify this code-switching phenomena, we have also conducted a case study (4.7.1) which shows that searching the information about translated entities online yields a much higher failure rate than searching them in their original languages. Motivated by these observations, we define *three unexplored use cases* of multilingual ToD where a foreign-language speaker uses ToD in the foreign-language country (**F&F**) or an English country (**F&E**), and an English speaker uses ToD in a foreign-language country (**E&F**). These use cases are different from the traditional **E&E** use case where an English speaker uses ToD in an English-speaking country.

To bridge the aforementioned gap between existing data curation methods and the real use cases, we propose a novel data curation method that *globalizes* an existing multi-domain ToD dataset beyond English for the three unexplored use cases.

Specifically, building on top of MultiWoZ [124] — an English ToD dataset for dialogue state tracking (DST), we create GlobalWoZ, a new multilingual ToD dataset in three new target-languages via machine translation and crawled ontologies in the target-language countries.

Our method only requires minor human efforts to post-edit a few hundred machine-translated dialogue templates in the target languages for evaluation. Besides, as cross-lingual transfer via pre-trained multilingual models [60, 93, 233, 234] has proven effective in many cross-lingual tasks, we further investigate another question: *How do these multilingual models trained on the English ToD dataset transfer knowledge to our globalized dataset?* To answer this question, we prepare a few baselines by evaluating popular ToD systems on our created test datasets in a *zero-shot* cross-lingual transfer setting as well as a *few-shot* setting.

Our contributions include the following:

- To the best of our knowledge, we provide the first step towards analyzing three unexplored use cases for multilingual ToD systems.
- We propose a cost-effective method that creates a new multilingual ToD dataset from an existing English dataset. Our dataset consists of high-quality test sets which are first translated by machines and then post-edited by professional translators in three target languages (Chinese, Spanish and Indonesian). We also leverage machine translation to extend the language coverage of test data to another 17 target languages.
- Our experiments show that current multilingual systems and translate-train methods fail in zero-shot cross-lingual transfer on the dialogue state tracking task. To tackle this problem, we propose several data augmentation methods to train strong baseline models in both zero-shot and few-shot cross-lingual transfer settings.

4.2 Related Work

Over the last few years, the success of ToD systems is largely driven by the joint advent of neural network models [221–223] and collections of large-scale annotation corpora. These corpora cover a wide range of topics from a single domain (e.g.,

ATIS [235], DSTC 2 [236], Frames [237], KVRET [221], WoZ 2.0 [238], M2M [239]) to multiple domains (e.g., MultiWoZ [124], SGD [240]). Most notably among these collections, MultiWoZ is a large-scale multi-domain dataset that focuses on transitions between different domains or scenarios in real conversations [124]. Due to the high cost of collecting task-oriented dialogues, only a few monolingual or bilingual non-English ToD datasets are available [225, 226, 241]. While there is an increasing interest in data curation for multilingual ToD systems, a vast majority of existing multilingual ToD datasets do not consider the real use cases when using a ToD system to search for local entities in a country. We fill this gap in this paper to provide the first analysis on three previously unexplored use cases.

4.3 Data Curation Methodology

In order to globalize an existing English ToD dataset for the three aforementioned use cases, we propose an approach consisting of four steps as shown in Figure 4.2: (1) we first extract dialogue templates from the English ToD dataset by replacing English-specific entities with a set of general-purpose placeholders (4.3.1); (2) we then translate the templates to a target language for both training and test data, with one key distinction that we only post-edit the test data by professional translators to ensure the data quality for evaluation (4.3.2); (3) next, we collect ontologies [242] containing the definitions of dialogue acts, local entities and their attributes in the target-language countries (4.3.3); (4) finally, we tailor the translated templates by automatically substituting the placeholders with entities in the extracted ontologies to construct data for the three use cases (4.3.4).

4.3.1 Automatic Template Creation

We start with MultiWoZ 2.2 [243] – a high-quality multi-domain English ToD dataset with more accurate human annotations compared to its predecessors MultiWoZ 2.0 [124] and MultiWoz 2.1 [244]. For the sake of reducing human efforts for collecting ToD context in the target languages, we re-use the ToD context written by human in MultiWoZ as the dialogue templates. Specifically as shown in Figure 4.2, we replace the English entities in MultiWoz by a set of general-purpose placeholders such as `[attraction-name0]` and `[attraction-postcode1]`, where

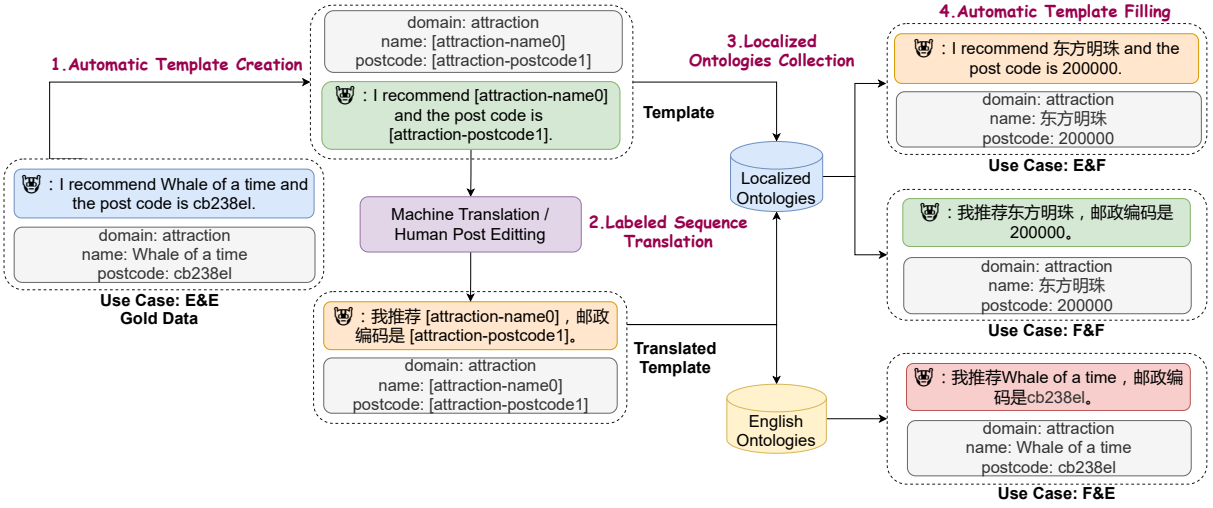


FIGURE 4.2: Illustration of our proposed pipeline: 1. Automatic Template Creation 2. Labeled Sequence Translation 3. Localized Ontologies Collection 4. Automatic Template Filling

each placeholder contains the entity's domain, attribute and ID. To do so, we first build a dictionary with entity-placeholder pairs by parsing the annotations of all dialogues. For example, from a dialogue text — *"I recommend Whale of a time and the post code is cb238el."*, we obtain two entity-placeholder pairs from its human annotations, i.e., (*Whale of a time*, `[attraction-name0]`) and (*cb238el*, `[attraction-postcode1]`). Next, we identify entities in the dialogue by their word index from the human annotations, replace them with their placeholders in the dictionary, and finally obtain dialogue templates with placeholders. Notably, we skip the entities with their attributes of `[choice]` and `[ref]` that represent the number of choices and booking reference number, as these attributes could be used globally.

4.3.2 Labeled Sequence Translation

Following [245] that translates sentences with placeholders, we use a machine translation system[1] to translate dialogue templates with our designed placeholders. As we observe, a placeholder containing an entity domain, attribute and ID (e.g., `attraction-name0`) is useful to provide contextually meaningful information to the translation system, thus usually resulting in a high-quality translation with

[1]We use Google Translate (`https://cloud.google.com/translate`), an off-the-shelf MT system.

the placeholder unchanged. This also enables us to easily locate the placeholders in the translation output and replace them with new entities in the target language.

To build a high-quality test set for evaluation, we further hire professional translators to post-edit a few hundred machine-translated templates, which produces natural and coherent sentences in the target languages. With the goal of selecting representative test templates for post-editing, we first calculate the frequency of all the 4-gram combinations in the MultiWoZ data, and then score each dialogue in the test set by the sum of the frequency of all the 4-gram combinations in the dialogue divided by the dialogue's word length. We use this scoring function to estimate the representiveness of a dialogue in the original dataset. Finally, we select the top 500 high-scoring dialogues in the test set for post-editing. We also use the same procedure to create a small high-quality training set for few-shot cross-lingual transfer setting.

4.3.3 Collection of Local Ontology

Meanwhile, we crawl the attribute information of local entities in three cities from public websites (e.g., tripadvisor.com, booking.com) to create three ontologies for the three corresponding target languages respectively. We select Barcelona for Spanish (an Indo-European language), Shanghai for Mandarin (a Sino-Tibetan language) and Jakarta for Indonesian (an Austronesian language), which cover a set of typologically different language families.

Given a translated dialogue template, we can easily sample a random set of entities for a domain of interest from a crawled ontology and assign the entities to the template's placeholders to obtain a new dialogue in the target language. Repeating this procedure on each dialogue template, we can easily build a high-quality labeled dataset in the target language. The number of our collected entities are either larger than or equal to those in the English data except for the "train" domain; we collected the information about only 100 "trains" for each languages due to the complexity in collecting relevant information.

4.3.4 Template Filling for Three Use Cases

After the above steps, we assign entities in a target language to the translated templates in the same target language for the F&F case, while assigning target-language entities to the English (source-language) templates for the F&E case. As for the E&F case, we keep the original English context by skipping the translation step and replace the placeholders with local entities in the target language (see Figure 4.2 for examples).

To sum up, our proposed method has three key properties: (1) our method is *cost-effective* as we only require a limited amount of post-editing efforts for a test set when compared to the expensive crowd-sourced efforts from the other studies; (2) we can easily sample entities from an ontology to create *large-scale machine-translated data* as a way of data augmentation for training; (3) our method is *flexible* to update entities in a ToD system whenever an update of ontology is available, e.g., extension of new entities.

4.4 Task & Settings

4.4.1 Dialogue State Tracking

Our experiments focus on the dialogue state tracking (DST), one of the fundamental components in a ToD system that predicts the goals of a user query in multi-turn conversations. We follow the setup in MultiWoZ [124] to evaluate ToD systems for DST by the joint goal accuracy which measures the percentage of correctly predicting all goals in a multi-turn conversation.

4.4.2 Experimental Settings

Zero-Shot Cross-lingual Transfer: Unlike prior studies that annotate a full set of high-quality training data for a target language, we investigate the *zero-shot cross-lingual transfer* setting where we have access to only a high-quality human-annotated English ToD data (referred to as gold standard data hereafter). In addition, we assume that we have access to a machine translation system that

translates from English to the target language. We investigate this setting to evaluate how a multilingual ToD system transfers knowledge from a high-resource source language to a low-resource target language.

Few-Shot Cross-lingual Transfer: We also investigate few-shot cross-lingual transfer, a more practical setting where we are given a small budget to annotate ToD data for training. Specifically, we include a small set (100 dialogues) of high-quality training data post-edited by professional translators (4.3.2) in a target language, and evaluate the efficiency of a multilingual ToD on learning from a few target-language training examples.

4.5 Proposed Baselines

We prepare a base model for GlobalWoZ in the zero-shot and few-shot cross-lingual transfer settings. We select Transformer-DST [246] as our base model as it is one of the state-of-the-art models on both MultiWoZ 2.0 and MultiWoZ 2.1[2]. In our paper, we replace its BERT encoder with an mBERT encoder [93] for our base model and propose a series of training methods for GlobalWoZ. As detailed below, we propose several data augmentation baselines that create different training and validation data for training a base model. Note that all the proposed baselines are model agnostic and the base model can be easily substituted with other popular models [223, 247]. For each baseline, we first train a base model on its training data for 20 epochs and use its validation set to select the best model during training. Finally we evaluate the best model of each baseline on the same test set from GlobalWoZ. We will release GlobalWoZ and our pre-trained models to encourage faster adaptation to future research.

4.5.1 Pure Zero-Shot (E&E)

We train a base model on the gold standard English data (E&E) and directly apply the learned model to the test data of the three use cases in GlobalWoZ. With this method, we simulate the condition of having labeled data only in the

[2]According to the leaderboards of Multi-domain Dialogue State Tracking on MultiWoZ 2.0 and MultiWoZ 2.1 on paperwithcode.com as of 11/15/2021.

source language for training, and evaluate how the model transfers knowledge from English to the three use cases. We use **Zero-Shot (E&E)** to denote this method.

4.5.2 Translate-Train

We use our data curation method (4.3) to translate the templates by an MT system but replace the placeholders in the translated templates with machine-translated entities to create a set of pseudo-labeled training data. Next, we train a base model on the translated training data without local entities, and evaluate the model on the three use cases.

We denote this method as **Translate-Train**.

4.5.3 Single-Use-Case Training

By skipping the human post-editing step in our data curation method (4.3), we leverage a machine translation system to automatically create a large set of pseudo-labeled training data with local entities for the three use cases. In the F&F case, we translate the English templates by the MT system and replace the placeholders in the translated templates with foreign-language entities to create a training dataset. In the F&E case, we replace the placeholders in the translated templates with the original English entities to create a code-switched training dataset. In the E&F case, we use the original English templates and replace the placeholders in the English templates with foreign-language entities to create a code-switch training dataset. With this data augmentation method, we can train a base model on each pseudo-labeled training dataset created for each use case. We denote this method as **SUC** (Single-Use-Case).

4.5.4 Bi-/Multi-lingual Bi-Use-Case Training

We investigate the performance of combining the existing English data and the pseudo-labeled training data created for one of the three use cases (i.e., F&F, F&E, E&F), one at a time, to do bi-use-case training. In the bilingual training, we only combine the gold English data (E&E) with the pseudo-labeled training data

in one target language in one use case for joint training. We denote this method as **BBUC** (Bilingual Bi-Use-Case). In the multilingual training, we combine gold English data (E&E) and pseudo-labeled training data in all languages in one use case for joint training. We denote this method as **MBUC** (Multilingual Bi-Use-Case).

4.5.5 Multilingual Multi-Use-Case Training

We also propose to combine the existing English data (E&E) and all the pseudo-labeled training data in all target languages for all the use cases (F&F, F&E, E&F). We then train a single model on this combined multilingual training dataset and evaluate the model on test data in all target languages for all three use cases. We denote this method as **MMUC** (Multilingual Multi-Use-Case).

4.6 Experiment Results

In this section, we show the results of all methods in the zero-shot (4.6.1) and few-shot (4.6.2) settings.

4.6.1 Zero-shot Cross-lingual Transfer

4.6.1.1 Use Case F&F, F&E and E&F

Table 4.1 reports the joint goal accuracy of all proposed methods on the three different sets of test data in the F&F, F&E, and E&F use cases. Both Zero-Shot (E&E) and Translate-Train struggle, achieving average accuracy of less than 10 in all use cases. Despite its poor performance, Zero-Shot (E&E) works much better in F&E than F&F, while its results in F&F and E&F are comparable, indicating that a zero-shot model trained in E&E can transfer knowledge about local English entities more effectively than knowledge about English context in downstream use cases. Besides, we also find that Zero-Shot (E&E) performs better on the Spanish or Indonesian context than the Chinese context in F&E. One possible reason is

Case	Methods	zh	es	id	avg
F&F	Zero-Shot (E&E)	1.22	1.38	1.26	1.28
	Translate-Train	2.61	2.59	5.74	3.65
	SUC (F&F)	36.97	24.66	25.26	28.96
	BBUC (E&E + F&F)	37.32	25.52	26.39	29.74
	MBUC (E&E + F&F)	**38.01**	**26.03**	**28.22**	**30.76**
F&E	Zero-Shot (E&E)	6.92	11.34	9.09	9.12
	Translate-Train	2.28	4.97	4.67	3.97
	SUC (F&E)	56.28	41.94	47.93	48.71
	BBUC (E&E + F&E)	59.87	48.20	54.79	54.29
	MBUC (E&E + F&E)	**60.37**	**53.56**	**54.93**	**56.28**
E&F	Zero-Shot (E&E)	1.69	1.81	1.82	1.77
	Translate-Train	1.39	1.76	1.86	1.67
	SUC (E&F)	38.56	28.00	43.82	36.79
	BBUC (E&E + E&F)	39.87	27.29	45.48	37.54
	MBUC (E&E + E&F)	**40.20**	**29.22**	**47.06**	**38.83**

TABLE 4.1: Zero-shot cross-lingual accuracy on DST over three target languages in three use cases.

that English is closer to the other Latin-script languages (Spanish and Indonesian) than Chinese.

Our proposed data augmentation methods (SUC, BBUC, MBUC) perform much better than non-adapted methods (Zero-Shot (E&E) and Translate-Train) that do not leverage any local entities for training. In particular, it is worth noting that even though Translate-Train and SUC both do training on foreign-language entities in F&F and E&F, there is a huge gap between these two methods, since Translate-Train has only access to the machine-translated entities rather than the real local entities used by SUC. This huge performance gaps not only show that Translate-Train is not an effective method in practical use cases but also prove that having access to local entities is a key to building a multilingual ToD system for practical usage.

Comparing our data augmentation methods SUC and BBUC, we find that the base model can benefit from training on additional English data (E&E), especially yielding a clear improvement of up to 5.58 average accuracy points in F&E. Moreover, when we increase the number of languages in the bi-use-case data augmentations (i.e., MBUC), we observe an improvement of around 1 average accuracy points in all three use cases w.r.t. BBUC. These observations encourage a potential future direction that explores better data augmentation methods to create high-quality pseudo-training data.

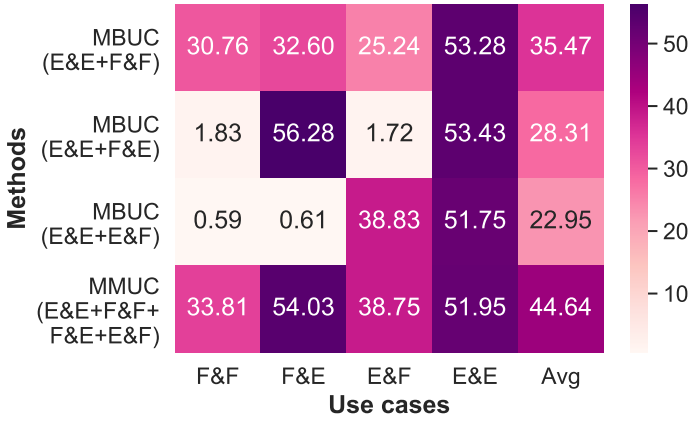


FIGURE 4.3: Performance of MMUC vs MBUC on the test data of the four use cases, F&F, F&E, E&F and E&E.

4.6.1.2 One Model for All

Notice that we can train a single model by MMUC for all use cases rather than training separate models, one for each use case. In Figure 4.3, we compare MMUC and MBUC (rows) on the test data in the four use cases (columns). Although MMUC may not achieve the best results in each use case, it achieves the best average result over the four use cases, indicating the potential of using one model to simultaneously handle all the four use cases.

4.6.2 Few-shot Cross-lingual Transfer

In few-shot experiments, we use the same scoring function based on frequency of all 4-gram combinations (4.3.2) to select 100 additional dialogues from train set for human-post editing, and create high-quality training data for each of the three use cases. To avoid overfitting on this small few-shot dataset, we combine the few-shot data with the existing English data for training a base model (Few-Shot+Zero-Shot (E&E)). Next, we also investigate a model trained with additional synthetic data created by our proposed SUC. In Figure 4.4, we find that our proposed SUC without additional few-shot data has already outperformed the model trained with few-shot data and English data (Few-shot + Zero-Shot (E&E)), indicating that the model benefit more from a large amount of pseudo-labeled data than a small set of human-labeled data. If we combine the data created by SUC with the few-shot data or with both few-shot and English data to train the model, we observe

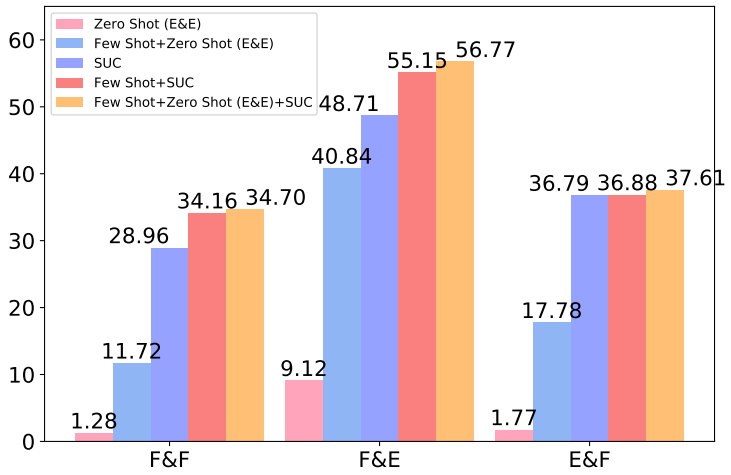


FIGURE 4.4: Few-shot cross-lingual average joint accuracy on DST over three target languages in three use cases.

improvements over SUC, especially with a clear gain of 8.06 accuracy points in F&E.

4.7 Discussion

4.7.1 Motivation for Code-Switched Use Cases

One key research question is to validate whether code-switched use cases with local entities (i.e., F&E, E&F) are practically more useful for information seeking. To answer this question, we compare the failure rate of using local entities and machine-translated entities in information search, which is a proxy to the efficiency of using these two types of entities in conversations. We first randomly select 100 entities (33 attractions, 33 hotels and 34 restaurants) of Cambridge, Shanghai, Barcelona and Jakarta. We translate the English entities into Mandarin, Spanish and Indonesian and the foreign-language entities into English via Google Translate. We then manually search the translated entities on Google to check whether we can find the right information of the original entities. Notice that the failure of the above verification partially come from the translation error made by Google Translate, or the search failure due to the fact that this entity does not have a bilingual version at all. In Table 4.2, we observe a high failure rate of around 60% for almost all translated directions (except Zh→En) due to translation and search failures, significantly exceeding the low failure rate of searching original entities

Translate	Search	En→Zh	En→Es	En→Id	Zh→En	Es→En	Id→En
✔	✔	35	42	36	62	30	31
✔	✗	61	34	51	18	18	15
✗	✔	0	24	13	11	50	54
✗	✗	4	0	0	8	2	0
Failure Case (MTed Entities)		65	58	64	37	70	69
Failure Rate (MTed Entities)		65%	58%	64%	37%	70%	69%
Failure Rate (Original Entities)		3%	3%	3%	0%	1%	0%

TABLE 4.2: The search and translation results of 100 translated entities on Google. En→Zh refers to the translation of English entities to Mandarin and Zh→En refers to the translation of Mandarin entities to English.

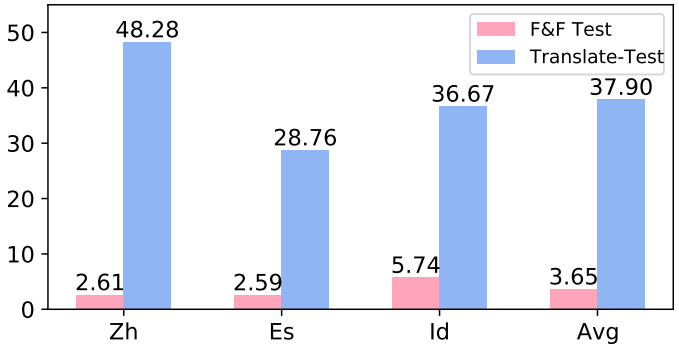


FIGURE 4.5: Joint accuracy of Translate-Train for DST on the F&F Test vs Translate-Test data.

online. Besides, even if we can find the right information of the translated entities, local people may not recognize or use the translated entities for communication, thus this results in inefficient communication with local people.

4.7.2 Overestimate of Translate-Train

In previous translation-based work, a multilingual ToD system is usually built based on the translation of English training data (Translate-Train), and is evaluated on translated test data without any local entities (Translate-Test). To verify whether this procedure is reliable to build a multilingual ToD system, we also create a test dataset with translated entities instead of local entities in the target languages. As shown in Figure 4.5, we find the Translate-Train model performs well on the test data with translated entities, but performs badly on the test data with real local entities. To the best of our knowledge, we provide the first analysis

to identify this performance gap between the translated test data and data with real local entities in a more realistic use case. Our work sheds light on the development of a globalized multilingual ToD system in practical use cases. We can tackle the challenge of localization issues by exploring new data augmentation method. Alternatively we can also explore new methods from the model level by building modular network to update the entities or perform transfer learning to adapt to new case without retraining.

4.7.3 Local Context vs. Local Entities

We compare the impact of training a model on data with either local contexts or local entities when the model is evaluated on monolingual test data in F&F and E&E. Specifically, when the train set has access to local context only, all the entities in the train set are replaced by entities in non-target languages. Similarly, when the train set has access to local entities only, the contexts in the train set are replaced by context in the non-target languages. Table 4.3 shows that both local contexts and local entities are essential to building ToD systems in the target language.

Train Set	**E&E (en)**	**F&F (zh)**	**F&F (es)**	**F&F (id)**	**avg**
Local Context Only	5.46	1.77	2.37	2.40	3.20
Local Entities Only	6.39	0.36	2.41	2.75	3.05
Local Context & Entities	**52.78**	**36.97**	**24.66**	**25.26**	**38.13**

TABLE 4.3: Comparison of training with local context or/and local entities on the joint accuracy for DST in E&E (en) and F&F (zh, es, id).

4.7.4 Scaling up to 20 Languages

With our proposed data curation method, it is possible to extend the dataset to cover more languages without spending extra costs if we skip the human post-editing step. Before doing so, one key question is whether the evaluation on the translated data without human post-editing is reliable as a proxy of the model performance. Thus, we conduct the experiments by evaluating the model performance of all baselines (4.5) on two sets of test data built with local entities: (1) **MT** test data where translated template is created by machine translation only (4.3.2); (2)

Use Case	**F2F**		**F2E**	
Methods	**MT Test**	**MTPE Test**	**MT Test**	**MTPE Test**
Zero-Shot (E&E)	1.29	1.28	9.64	9.12
Translate-Train	3.71	3.65	4.17	3.97
SUC	35.78	28.96	56.15	48.71
BBUC	36.31	29.74	57.84	54.29
MBUC	**37.89**	**30.76**	**58.76**	**56.28**
Spearman's correlation	**1.0**		**1.0**	

TABLE 4.4: Comparison of average joint accuracy on DST reported on MT test data and MTPE test data for use case F&F and F&E

MTPE test data where translated template is first translated by machines and post-edited later by professional translators. As shown in Table 4.4, the overall reported results on MT test data are higher than those reported on MTPE test data, which is expected because the distribution of the MT test data is more similar to the MT training data. Although there are some differences on individual languages, the conclusions derived from the evaluations on the MT test data remain the same as those derived from the evaluation on the MTPE test data. We also calculate the Spearman rank correlation coefficient between the average results reported on MTPE test data and MT test data in Table 4.4, which shows a statistically high correlation between the system performance on the MT test data and MTPE test data. Therefore, we show that the MT test data can be used as a proxy to estimate the model performance on the real test data for more languages. Thus we build MT test data for another 17 languages that are supported by Google Translate, Trip Advisor and Booking.com at the same time. Table 4.5 shows the results of Zero-Shot (E&E) and SUC on the test data of F&F, F&E and E&F in 20 languages. The results show that the model has the best performance in the F&E use case compared with the other two use cases, which is consistent with our findings in Table 4.1.

4.8 Chapter Summary

In this chapter, we embark on an in-depth exploration of three novel and hitherto unexplored applications for multilingual task-oriented dialogue (ToD) systems. Our

Case	Method	Avg
F&F	Zero-Shot (E&E)	1.48
	SUC	**16.12**
F&E	Zero-Shot (E&E)	9.03
	SUC	**34.20**
E&F	Zero-Shot (E&E)	1.97
	SUC	**23.40**

TABLE 4.5: Average results of Zero-Shot (E&E) on test data of F&F, F&E and E&F in 20 languages.

investigation is anchored around the formulation and implementation of an innovative approach to data curation. This approach distinctively integrates a machine translation system with local entities specific to the target languages, thereby facilitating the creation of an unprecedented multilingual ToD dataset, herein referred to as GlobalWoZ.

Elaborating further, we introduce a comprehensive set of robust baseline methodologies. These methodologies have been meticulously designed and are instrumental in laying the groundwork for subsequent research endeavors in the realm of multilingual ToD systems. To validate the effectiveness and applicability of these methodologies, we have conducted a series of extensive and rigorous experiments utilizing the GlobalWoZ dataset. These experiments are pivotal in demonstrating the practicality and potential of the proposed baseline methods.

Moreover, a significant aspect of our research is the expansion of linguistic coverage within the domain of multilingual ToD systems. We have successfully extended this coverage to encompass a total of 20 languages. This expansion is not just a mere quantitative increase; it represents a substantial stride towards the realization of a truly globalized multilingual ToD system. Such a system is envisioned to serve the diverse linguistic needs of the world's populace, bridging communication gaps and fostering more inclusive and effective dialogue across a myriad of languages.

In conclusion, the contributions of this chapter are multifaceted. Firstly, it presents a novel methodology for data curation that is poised to revolutionize the way multilingual ToD datasets are developed. Secondly, it establishes a set of strong baseline methods that lay a solid foundation for future research in this field. Thirdly, through exhaustive experimental analysis, it validates the effectiveness of these

methods. Finally, it significantly broadens the linguistic scope of multilingual ToD systems, taking a decisive step towards the creation of a globalized platform that caters to the linguistic diversity of the global citizenry.

4.9 Ethical Review

In this section, we would like to address the ethical concerns. All the professional translators in this project have been properly compensated. For Chinese and Spanish, we have followed the standard procurement requirements and engaged three translation companies for quality and price comparison. A small sample of the data had been given to them for MTPE and we then compared their translation results. Following that, we selected the company that produced the best sample translation, and submitted the full translation orders according to the agreed price quotations. For Indonesian, three translation companies were also requested to provide sample MTPE, but our quality check found the quality of these samples to be unsatisfactory. So, no company was engaged, and our in-house Indonesian linguistic resources were used instead. These Indonesian linguists were assigned to work on this project during normal working hours and given proper compensation complying with the local labor laws.

4.10 Supplementary Material

4.10.1 Comparison of Four Use Cases

Use Case	Source ToD	Speaker (ToD Context)	Country (ToD Ontology)
F&F	English	Foreign Lang.	Foreign Lang.
F&E		Foregin Lang.	English
E&F		English	Foreign Lang.
E&E		English	English

TABLE 4.6: Four use cases of multilingual ToD systems: A foreign language or English speaker travels to a country of a foreign language or English.

4.10.2 Examples of Labeled Sequence Translation

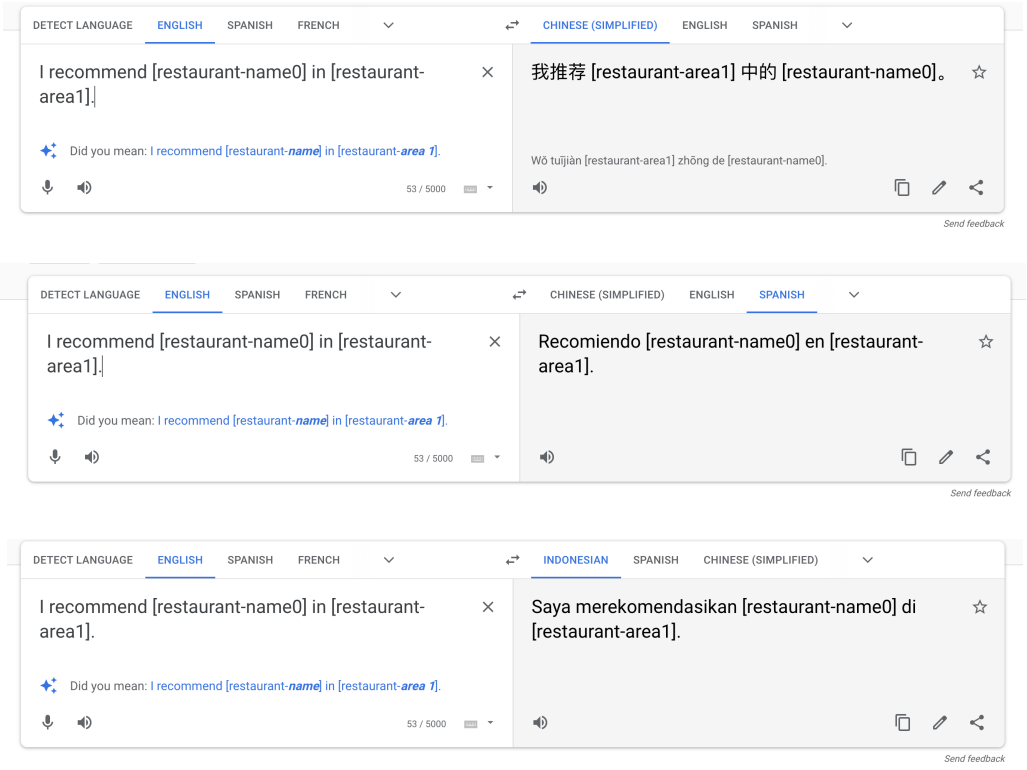


FIGURE 4.6: An instance of labeled sequence translation with google translate, from English to three target languages, Mandarin, Spanish and Indonesian.

4.10.3 BLEU Score of MT versus MTPE Test Template

Languages	Zh	Es	Id	Avg
BLEU Score	55.61	49.33	48.97	51.30

TABLE 4.7: BLEU Scores of MT Test Template using MTPE Test Template as reference.

4.10.4 Test Set Distribution

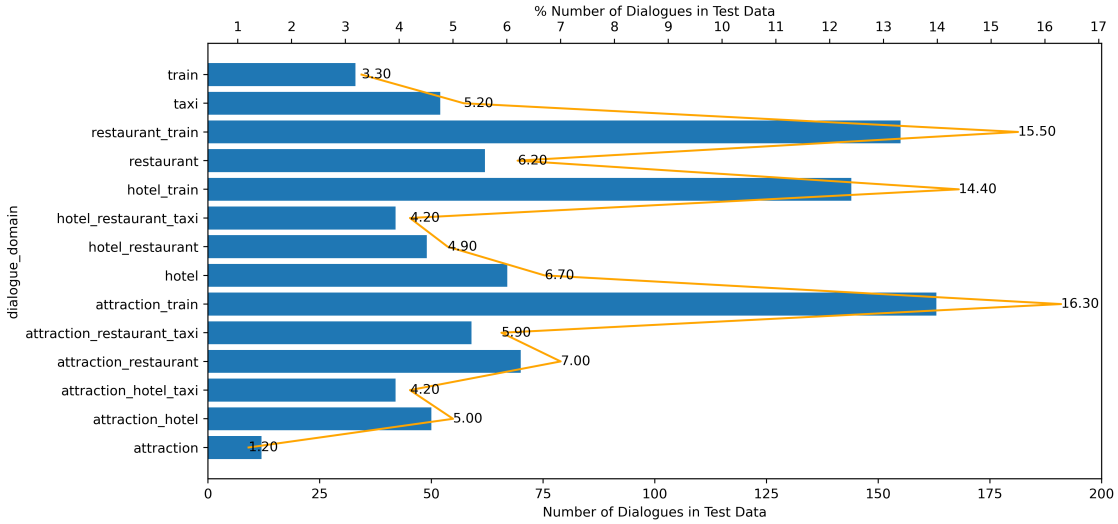


FIGURE 4.7: Gold English Test Set Distribution by Domains. We follow this distribution to select the top 500 high-scoring dialogues in the test set for post-editing.

4.10.5 Selected Languages

Language	ISO639-1code	Language Family	# Wikipedia articles (in millions)	High / Middle/ Low Resource	Writing Script	Selected City
English	en	IE: Germanic	6.35	High	Latin	Cambridge
Swedish	sv	IE: Germanic	2.95	High	Latin	Stockholm
German	de	IE: Germanic	2.61	High	Latin	Berlin
French	fr	IE: Romance	2.35	High	Latin	Paris
Dutch	nl	IE: Germanic	2.06	High	Latin	Amsterdam
Russian	ru	IE: Slavic	1.74	High	Cyrillic	Moscow
Italian	it	IE: Romance	1.71	High	Latin	Rome
Spanish	es	IE: Romance	1.71	High	Latin	Barcelona
Japanese	ja	Japonic	1.28	High	Ideograms	Tokyo
Vietnamese	vi	Austro-Asiatic	1.27	High	Latin	Ho Chi Minh City
Mandarin	zh	Sino-Tibetan	1.22	High	Chinese ideograms	Shanghai
Arabic	ar	Afro-Asiatic	1.13	High	Arabic	Cairo
Portuguese	pt	IE: Romance	1.07	High	Latin	Lisbon
Indonesian	id	Austronesian	0.59	Middle	Latin	Jakarta
Norwegian	no	IE: Germanic	0.56	Middle	Latin	Oslo
Korean	ko	Koreanic	0.55	Middle	Hangul	Seoul
Turkish	tr	Turkic	0.42	Middle	Latin	İstanbul
Hebrew	he	Afro-Asiatic	0.30	Low	Hebrew	Tel Aviv
Danish	da	IE: Germanic	0.27	Low	Latin	Copenhagen
Greek	el	IE: Greek	0.20	Low	Greek	Athens
Thai	th	Kra-Dai	0.14	Low	Brahmic	Bangkok

TABLE 4.8: Statistics about languages in the cross-lingual benchmark. The selected 21 languages (including English) belong to 8 language families and 1 isolate, with Indo-European (IE) having the most members. We categorize the languages with more than 1 million, more than 400 thousand but less than 1 million, less than 400 thousand Wikipedia articles as high resource languages, middle resource languages and low resource languages. For each language, we select one city for each language to collect localized ontology.

4.10.6 Statistics of Entities in the Collected Ontology

Languages	rest.	hotel	attr.	train	taxi
en	110	33	79	2828	222
zh	3000	496	1000	100	4496
es	3000	426	1000	100	4426
id	3000	999	792	100	4791
ar	2989	680	1000	100	4669
da	2343	165	1000	100	3508
de	2988	659	1000	100	4647
el	2600	1000	1000	100	4600
fr	3000	1000	1000	100	5000
he	1558	258	1000	100	2258
it	3000	800	1000	100	2800
ja	2967	864	1000	100	4831
ko	2990	532	1000	100	4522
nl	2990	537	1000	100	4527
no	1293	95	757	100	2145
pt	2993	951	1000	100	4944
ru	2985	531	1000	100	4516
sv	3000	214	891	100	4105
th	2995	1000	1000	100	4995
tr	2986	533	1000	100	4519
vi	2991	773	1000	100	4764

TABLE 4.9: Statistics of entities in the collected ontology in different languages. We count the number of entities in the database of each domain. Noticed that in the Taxi database of MultiWoZ, it only list down the taxi colors, taxi types and taxi phones. The taxi destination and departure refer to the entities in the restaurant, hotel and attraction domains. Thus, we use the sum of the number of entities in Restaurant, Hotel and Attraction domains as a proxy of the total number of entities in taxi domain. Besides, we follow MultiWoZ to collect one hospital and one police station for each city.

4.10.7 Statistics of GlobalWoZ

Use Case	F&F				F&E				E&F			
Languages	Train & Dev	Method	Test	Method	Train & Dev	Method	Test	Method	Train & Dev	Method	Test	Method
zh	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
es	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
id	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
ar	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
da	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
de	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
el	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
fr	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
he	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
it	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ja	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ko	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
nl	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
no	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
pt	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ru	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
sv	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
th	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
tr	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
vi	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human

TABLE 4.10: Statistics of created dataset, GlobalWoZ for each use case in each target language. For E&F, as the context is the original Engish data, we consider it is created by human. For test data of zh, es and id, we replace the entities twice to boostrap the test data to 1000 dialogues. We are currently preparing the post editing of the other 500 dialogues in test data. Meanwhile, we are leveraging machine translation to prepare the train data for the 17 languages and will release it with baselines in the next version soon.

4.10.8 Dialogue Examples

FIGURE 4.8: Examples of some utterances in original E&E data, MT data and MTPE data,

4.10.9 Summary of Proposed Baselines

Methods	En Context	En Entities	Local Context	Local Entities	Translated Entites
Zero-Shot (E&E)	✔	✔			
Translate-Train			✔		✔
SUC (F&F)			✔	✔	
SUC (F&E)		✔	✔		
SUC (E&F)	✔			✔	

TABLE 4.11: Accessibility of different types of context and entities for each method.

Methods	E&E	F&F	F&E	E&F
Zero-Shot (E&E)	✔			
Translate-Train				
SUC (F&F)		✔		
SUC (F&E)			✔	
SUC (E&F)				✔
BBUC (E&E + F&F)	✔	✔		
BBUC (E&E + F&E)	✔		✔	
BBUC (E&E + E&F)	✔			✔
MBUC (E&E + F&F)	✔	✔		
MBUC (E&E + F&E)	✔		✔	
MBUC (E&E + E&F)	✔			✔
MMUC (E&E + F&F + F&E + E&F)	✔	✔	✔	✔

TABLE 4.12: Accessibility of data in each use case for each method. Noticed that Translate-Train doesn't have access to the data of the four use cases. Translate-Train has access to a set of pseudo-labeled training data created by replacing the placeholders in the translated template with machine-translated entities instead of local entities.

4.10.10 Use Case E&E

We also compare the performance of all methods on the original E&E test data. As **Zero-Shot (E&E)** is trained on monolingual English training data, it gets a high accuracy of 52.78 on the English test data. In contrast, **Translate-Train** and **SUC** (F&F) perform poorly on the English test data, because both of them have no access to any English data. Comparing to **SUC** (F&F), **SUC** (F&E) and **SUC** (E&F) achieve higher accuracy scores as they either have access to English context or English entities. When we perform bilingual and multilingual joint training (i.e., **BBUC** and **MBUC**), the base model has a performance increase except **MBUC** (E&E + E&F). This shows that bilingual and multilingual joint training may be used to improve the performance on source language. Further research can be done in this line.

Methods	**En**
Zero-Shot (E&E)	52.78
Translate-Train	2.27
SUC (F&F)	1.09
SUC (F&E)	6.39
SUC (E&F)	5.46
BBUC (E&E + F&F)	52.87
BBUC (E&E + F&E)	**53.69**
BBUC (E&E + E&F)	53.05
MBUC (E&E + F&F)	53.28
MBUC (E&E + F&E)	53.43
MBUC (E&E + E&F)	51.75

TABLE 4.13: Joint accuracy on DST in three target languages on the English test data.

4.10.11 Breakdown of Few Shot Results

Zero Shot (E&E)				
Use Case	Zh	Es	Id	Avg
F2F	1.22	1.38	1.26	1.28
F2E	6.92	11.34	9.09	9.12
E2F	1.69	1.81	1.82	1.77
Few Shot + Zero Shot (E&E)				
Use Case	Zh	Es	Id	Avg
F2F	15.93	7.13	12.09	11.72
F2E	39.88	39.38	43.26	40.84
E2F	20.61	14.17	18.55	17.78
SUC				
Use Case	Zh	Es	Id	Avg
F2F	36.97	24.66	25.26	28.96
F2E	56.28	41.94	47.93	48.71
E2F	38.56	28.00	43.82	36.79
Few Shot + SUC				
Use Case	Zh	Es	Id	Avg
F2F	37.81	25.15	39.51	34.16
F2E	58.39	53.03	54.02	55.15
E2F	38.75	27.66	44.23	36.88
Few Shot + Zero Shot (E&E) + SUC				
Use Case	Zh	Es	Id	Avg
F2F	37.52	26.44	40.15	34.70
F2E	59.21	54.93	56.17	56.77
E2F	39.51	27.84	45.48	37.61

TABLE 4.14: A breakdown of few-shot cross-lingual average joint accuracy on DST over three target languages in three use cases.

4.10.12 Concrete Examples where Translate-Train Performs Badly on the Test Data with Real Local Entities.

Through investigation, we found that the Translate-Train method usually performed badly in two main scenarios. Figure 5.5 is the illustrations of the two scenarios. Scenario 1 is when the Translate-Train can predict values that are close to the meaning of the ground truth values but suffer from the problems of translationese. For example, model trained with Translate-Train may predict "美食酒吧" (gastropub), which is a direct translation of gastropub and not commonly used in Chinese instead of "酒吧餐" (bar). Scenario 2 is when Translate-Train needs to predict the name of real localized entities which Translate-Train doesn't have access to. For example, trained with Translate-Train may predict "冈维尔酒店" (Gonville Hotel) which is a direct translation of Gonville Hotel, instead of "汉庭酒店" (Hanting Hotel) which is unseen in Translate-Train training data.

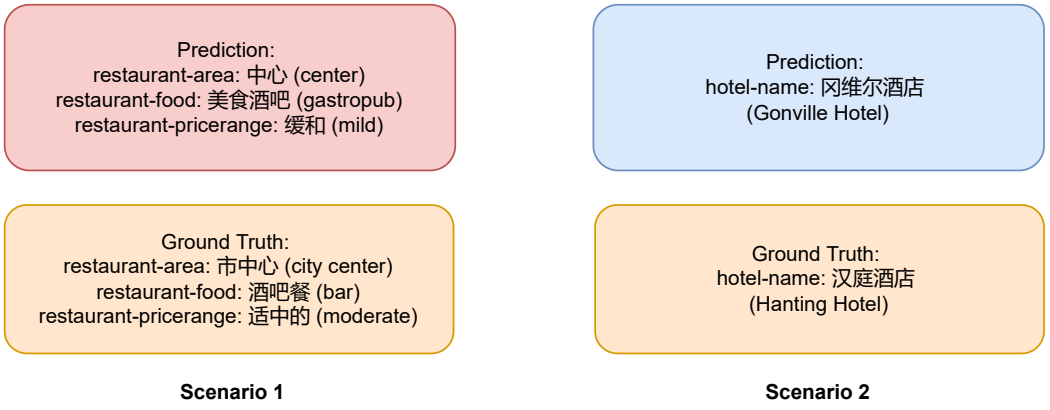


FIGURE 4.9: Concrete examples where Translate-Train performs badly on the test data with real local entities.

4.10.13 Breakdown of the Results of Local Context vs Local Entities by Languages

E&E (en)				
Context vs Entities	lavender Zh	Es	Id	Avg
En_Context	lavender 5.37	5.33	5.67	5.46
En_Entites	lavender 3.49	7.78	7.90	6.39
F&F (zh)				
Context vs Entities	lavender En	lavender Es	lavender Id	Avg
Zh_Context	lavender 1.74	lavender 1.77	lavender 1.80	1.77
Zh_Entites	lavender 0.27	lavender 0.73	lavender 0.10	0.36
F&F (es)				
Context vs Entities	En	lavender Zh	Id	Avg
Es_Context	1.73	lavender 2.01	3.37	2.37
Es_Entites	3.92	lavender 0.44	2.86	2.41
F&F (id)				
Context vs Entities	En	lavender Zh	Es	Avg
Id_Context	2.07	lavender 2.18	2.94	2.40
Id_Entites	3.92	lavender 0.84	3.48	2.75

TABLE 4.15: A breakdown of comparison of the impact of local context and local entities on joint accuracy for DST in each language. The cases where context and entities are in different script types are highlighted in lavender color.

Train Set	**different script type**	**same script type**
Local Context Only	**2.48**	3.52
Local Entities Only	0.98	**4.98**

TABLE 4.16: Comparison of the impact of script type on Local Context Only vs Local Entities Only. It shows that training with local entities is more important if the entities and contexts are written in the same type of language script (e.g. Latin script), otherwise training with local contexts is more important.

4.10.14 Breakdown of MT Test Data vs MTPE Test Data by Languages

Languages	Zh		Es		Id	
F2F	MT	MTPE	MT	MTPE	MT	MTPE
Zero-Shot (E&E)	1.19	1.22	1.40	1.38	1.28	1.26
Translate-Train	2.50	2.61	2.81	2.59	5.81	5.74
SUC	37.79	36.97	26.95	24.66	42.59	25.26
BBUC	38.62	37.32	27.34	25.52	42.96	26.39
MBUC	39.11	38.01	29.17	26.03	45.39	28.22
Spearman's correlation	1.00		1.00		1.00	
F2E	MT	MTPE	MT	MTPE	MT	MTPE
Zero-Shot (E&E)	7.61	6.92	11.67	11.34	9.64	9.09
Translate-Train	2.25	2.28	5.25	4.97	5.03	4.67
SUC	57.10	56.28	55.70	41.94	55.64	47.93
BBUC	59.05	59.87	57.68	48.20	56.80	54.79
MBUC	60.48	60.37	57.04	53.56	58.23	54.93
Spearman's correlation	1.00		0.90		1.00	

TABLE 4.17: Spearman rank correlation coefficient between the results on MTPE test data and MT test data for each language.

Chapter 5

Utilizing Large Language Models for Data Labeling in Natural Language Processing Tasks

5.1 Background

In the contemporary landscape of Large Language Models (LLMs), a paradigmatic shift in data annotation practices is observable, heralding a new era where the benefits of artificial intelligence become increasingly ubiquitous. This transition is marked by the move from traditional, labor-intensive annotation methods to more dynamic, AI-driven approaches, facilitating a significant increase in efficiency and accuracy. Large Language Models, by virtue of their extensive training on diverse datasets, are now equipped to understand and process information with a level of sophistication that closely mimics human cognitive abilities. Consequently, this evolution in data annotation not only accelerates the pace of data processing but also enhances the quality of insights derived from such data. As a result, the democratization of AI benefits is anticipated, with a broad spectrum of sectors – from healthcare and education to finance and transportation – poised to experience transformative impacts. This shift is not merely a technological advancement but a cornerstone in the journey towards a more interconnected and intelligent digital ecosystem, where the boundaries of AI's potential are continually expanding to encompass and enrich every aspect of human life.

The democratization of artificial intelligence (AI) [248, 249] aims to provide access to AI technologies to all members of society, including individuals, small- and medium-sized enterprises (SMEs), academic research labs, and nonprofit organizations. Achieving this goal is crucial for the promotion of innovation, economic growth, and fairness and equality. As typical AI models are usually data-hungry, one significant obstacle of AI democratization is the preparation of well-annotated data for training AI models.

Specifically, supervised learning critically depends on sufficient training data with accurate annotation, but data annotation can be a costly endeavor, particularly for small-scale companies and organizations [250]. The cost of data annotation typically includes the labor costs associated with the labeling process, as well as the time and resources required to hire, train and manage annotators. Additionally, there may be costs associated with the annotation tools and infrastructure needed to support the annotation process. Individuals or small-scale organizations may not have resources to annotate sufficient training data, thereby are unable to reap the benefits of contemporary AI technologies. Although the development of pre-trained language models such as BERT [251], XLNet [252], GPT-2 [94] and RoBERTa [111] eases the data-hungry issue to some extent, data annotation remains an unavoidable challenge for supervised model training.

GPT-3 [96, 253][1] is a powerful large language model developed by OpenAI. Evaluations show that GPT-3 has gained through pretraining a surprisingly wide range of knowledge, which can be transferred to downstream tasks through knowledge distillation [254]. Due to the model architecture and pretraining tasks designed for auto-regressive generation, GPT-3 is capable of generating human-like text and performing a broad array of NLP tasks, such as machine translation, summarization, and question-answering. However, the direct use of GPT-3 for inference in a production setting remains challenging due to its size and computational requirements. Moreover, such large language models often lack the flexibility of local deployment, since their parameters are usually not publicly available. In contrast, it is often more feasible to use smaller language model models, such as $\text{BERT}_{\text{BASE}}$ [251], in production environments.

In this chapter, we investigate the ability of GPT-3 to annotate training data for training machine learning models, which can substantially lower the annotation

[1]For brevity, we refer to both the original GPT-3 and InstructGPT as GPT-3.

cost and level the playing field for individuals or small organizations, so that they can harness the power of AI in their own missions. The process can be considered as distilling the knowledge of GPT-3 to small networks that can be straightforwardly deployed in production environments.

We conduct extensive experiments to evaluate the performance, time, and cost-effectiveness of 3 different GPT-3 based data annotation approaches for both sequence- and token-level NLP tasks. Our main contributions can be summarized as follows:

- We conduct comprehensive analysis of the feasibility of leveraging GPT-3 for data annotation for complex NLP tasks.
- We study 3 different GPT-3 based data annotation approaches, and then conduct extensive experiments on both sequence- and token-level NLP tasks to evaluate their performance.
- We find that directly annotating unlabeled data is suitable for tasks with small label space while generation-based methods are more suitable for tasks with large label space.
- We find that generation-based approaches tend to be more cost-effective compared with directly annotating unlabeled data.

5.2 Related Work

Large Language Models Large Language Models (LLMs) have made significant progress on natural language processing tasks in recent years. These models are trained with self-supervision on large, general corpora and demonstrate excellent performance on numerous tasks [96, 255–261]. LLMs possess the ability to learn in context through few-shot learning [96, 262]. Their capabilities expand with scale, and recent research has highlighted their ability to reason at larger scales with an appropriate prompting strategy [260, 263–267].

[268] investigate methods to utilize GPT-3 to annotate unlabeled data. However, they mainly focus on the generation and sequence classification tasks. In this work, we conduct more comprehensive experiments and analysis on a wider range of settings, covering both sequence- and token-level tasks. In a recent work, [269]

demonstrate a worker-and-AI collaborative approach for dataset creation with a few seed examples, while we also analyse approaches that support zero-shot training data generation, which do not require any seed examples.

Prompt-Learning Prompt-Learning, also known as Prompting, offers insight into what the future of NLP may look like [263, 265, 270]. By mimicking the process of pre-training, prompt-learning intuitively connects pre-training and model tuning [271]. In practice, this paradigm has proven remarkably effective in low-data regimes [272, 273]. For instance, with an appropriate template, zero-shot prompt-learning can even outperform 32-shot fine-tuning [274]. Another promising characteristic of prompt-learning is its potential to stimulate large-scale pre-trained language models (PLMs). When applied to a 10B model, optimizing prompts alone (while keeping the parameters of the model fixed) can yield comparable performance to full parameter fine-tuning [263]. These practical studies suggest that prompts can be used to more effectively and efficiently extract knowledge from PLMs, leading to a deeper understanding of the underlying principles of their mechanisms.

Data Augmentation There has been a significant amount of research in NLP on learning with limited labeled data for various tasks, including unsupervised pre-training [57, 251, 252, 275, 276], multi-task learning [277, 278], semi-supervised learning [279], and few-shot learning [280–282]. One approach to address the need for labeled data is through data augmentation [283, 284], which involves generating new data by modifying existing data points using transformations based on prior knowledge about the problem's structure [285]. The augmented data can be generated from labeled data [245, 286] and used directly in supervised learning [287] or employed in semi-supervised learning for unlabeled data through consistency regularization [288].

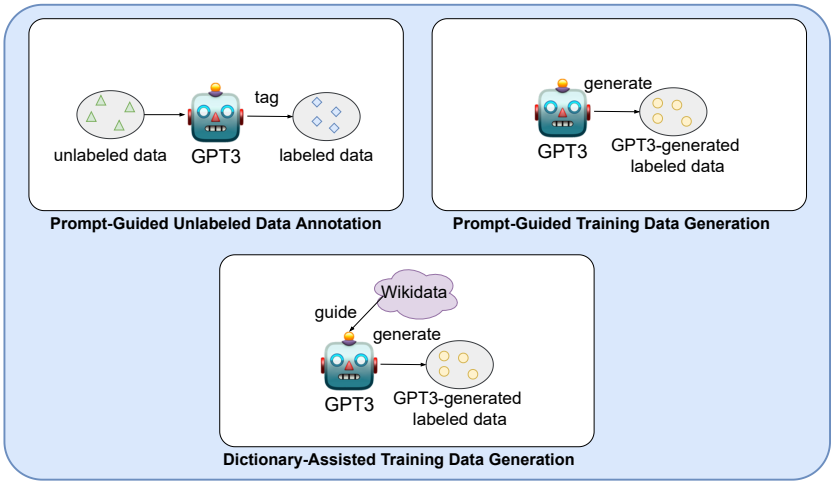


FIGURE 5.1: Illustrations of our proposed methods.

5.3 Methodology

We study 3 different approaches to utilize GPT-3 for data annotation: 1) prompt-guided unlabeled data annotation (PGDA); 2) prompt-guided training data generation (PGDG); and 3) dictionary-assisted training data generation (DADG). Illustrations are shown in Figure 5.1. Overall, these 3 approaches can be regarded as in-context learning [289], a new paradigm that is getting popular in NLP. Under this paradigm, a language model "learns" to do a task simply by conditioning on l_{IOP}, a list of input-output pairs (IOP). [2] More formally,

$$y_i = \text{GPT-3}(l_{\text{IOP}}, x_i) \tag{5.1}$$

where x_i is the query input sequence and y_i is the text generated by GPT-3. For comparison, the performance, cost, and time spent on the three methods are monitored. We also report the results of **Prompted Direct Inference (PGI)**, which is to instruct GPT-3 to directly annotate the test data.

5.3.1 Prompt-Guided Unlabeled Data Annotation (PGDA)

The first approach involves the creation of prompts to guide GPT-3 in annotating unlabeled data. To this end, task-specific prompts are designed to elicit labels

[2] Under the zero-shot settings, where l_{IOP} is not provided, our methods become instruction-tuning [290].

Choose the sentiment of the given text from Positive and Negative.
Text: a feast for the eyes
Sentiment: Positive
...
Text: boring and obvious
Sentiment: Negative
Text: [Unlabeled Data]
Sentiment: [Label]

FIGURE 5.2: An example of Prompt-Guided Unlabeled Data Annotation (PGDA) for SST2.

from GPT-3 for a given set of unlabeled data. In our experiments, the unlabeled data is derived from human-labeled datasets by removing the existing labels. The resulting GPT-3-labeled data is then used to train a local model to predict human-labeled test data, with the performance of this model being evaluated. As shown in Figure 5.2, an instruction with few-shot examples is given to GPT-3, followed by unlabeled data. GPT-3 is then prompted to predict labels for the unlabeled data.

5.3.2 Prompt-Guided Training Data Generation (PGDG)

The second approach is to utilize GPT-3 to autonomously generate labeled data for the specified task. This method involves the creation of prompts that guide GPT-3 to self-generate labeled data, which is subsequently used to train a local model to predict on human-labeled test data for the purpose of evaluation. For example, to generate training data with the relation "head of government", we can first "teach" GPT-3 to generate head-tail entity pairs that have the specified relation as illustrated in Figure 5.3. After we obtain the generated triplets (head-tail entity pairs with specified relation), as shown in Figure 5.4, we can then instruct GPT-3 to generate a sentence with the given entities and relation. Compared with tagging approach, a significant benefit of the generation-based approach is that it does not require a long list of label definitions specified in the prompt. For

Generate 20 different Head Entity and Tail Entity with the given Relation.
Relation: head of government
Relation Definition: head of the executive power of this town, city, municipality, state, country, or other governmental body
Relation: head of government
Head Entity: United States; **Tail Entity:** Chester Alan Arthur
...
Head Entity: Entity1; **Tail Entity:** Entity2

FIGURE 5.3: An example of prompting GPT-3 to generate entities for the relation "head of government" for FewRel.

Generate a sentence with the given entities and relation.
Relation: head of government
Head Entity: United States; **Tail Entity:** Chester Alan Arthur
Text: Chester Alan Arthur , 21st President of the United States , died of this disease , November 18 , 1886
...
Relation: head of government
Head Entity: Entity1; **Tail Entity:** Entity2
Text: [Generated Sentence]

FIGURE 5.4: An example of prompting GPT-3 to generate a sentence with the given entities and the relation "head of government" for FewRel.

example, to generate NER data, it can first generate entities of each entity type (e.g. organisation, person, etc.) and then generate a sentence with mixed entities.

5.3.3 Dictionary-Assisted Training Data Generation (DADG)

The third method is designed to utilize dictionary as an external source of knowledge to assist GPT-3 to generate labeled data for a specific domain. In our experiments, we choose Wikidata[3] as the dictionary. The data generated through this Wikidata-guided process is subsequently used to train a local model to predict human-labeled test data for the purpose of evaluating performance. For instance, to generate training data with the relation "head of government", we first query the head-tail entity pairs under the relation *P6*, relation ID of "head of government", from Wikidata. Upon obtaining the entity pairs from Wikidata, GPT-3 can then be instructed to generate a sentence with the specified entity pairs and relation. An advantage of this approach is that it can leverage knowledge base in specific domains, particularly when the domains are not present in the pre-trained corpus, thus allowing for the incorporation of external knowledge into GPT-3 without the need for fine-tuning.

5.4 Experiments

5.4.1 Experiment Settings

In this study, we conduct extensive experiments on both sequence- and token-level NLP tasks. The sequence-level tasks include sentiment analysis (SA) and relation extraction (RE). The token-level tasks include named entity recognition (NER) and aspect sentiment triplet extraction (ASTE).

More specifically, we use the SST2 dataset [113] for sentiment analysis, a well-known dataset comprising movie reviews. For relation extraction, we use FewRel [291], a large-scale relation extraction dataset. For NER, we use the AI domain split from the CrossNER dataset [292], which is the most difficult domain within the dataset and more closely mirrors real-world scenarios with its 14 entity types. For aspect sentiment triplet extraction, we use the laptop domain split released by [122].

[3] `https://www.wikidata.org`

To simulate the production scenario, we assume that the user has access to the off-shelf GPT-3 API. In all our experiments, we use *text-davinci-003*[4], the latest GPT-3 model. In addition, we assume that the user uses $\text{BERT}_{\text{BASE}}$ for production and has access to a few data points and Wikidata for each task. For each task, the resulting data of each approach is post-processed and reformatted into the same format of human-labeled data before being used to fine-tune a $\text{BERT}_{\text{BASE}}$ model. In order to accurately determine the cost and time required for human labeling, we conduct interviews and consultations with linguists and professional data annotators to obtain a precise estimation.

5.4.2 Sequence-Level Task

5.4.2.1 SST2

Annotation Approaches In PGDA, we randomly sample 10-shot data of the train set of the SST2 dataset to construct a prompt template, as illustrated in Figure 5.2. The prompt is used to guide GPT-3 in generating sentiment labels for the unlabeled data. In DADG, the ability of GPT-3 to perform Wikidata-guided few-shot generation is tested. We query entities in Wikidata from the movie domain. We then use the entities together with the same 10-shot data to prompt GPT-3 to generate sentences with a specified sentiment.

Results Table 5.1 presents the results of three different approaches. Overall, PGDA demonstrates the best performance among the three approaches. By labeling the same 3,000 data points, PGDA achieves an accuracy of 87.75, which is only 0.72 lower than that of human-labeled data. However, the cost and time consumed for PGDA are significantly lower than those for human labeling. By labeling 6,000 data, PGDA achieves a better performance than the human-labeled 3,000 data, while the cost is approximately 10% of the cost of human labeling. PGDG performes much worse than PGDA and human-labeled data. However, it also demonstrates a distinct advantage in terms of cost and time efficiency when generating the same amount of data compared with alternative approaches. DADG approach, which involves generating data with in-domain entities, does not result

[4]Released on 28 Nov 2022. Please refer to `https://beta.openai.com/docs/models` for more details.

Approach	Num. of Samples	Cost (USD)	Time (Mins)	Results
PGDA	3000	11.31	14†	87.75
	6000	22.63	27†	**89.29**
PGDG	3000	0.91	4†	73.81
	6000	1.83	8†	76.55
DADG	3000	7.18	23†	68.04
	6000	14.37	46†	71.51
Human Labeled	palepink 3000	221 - 300	1000	88.47
	palepink 67349	4800 - 6700	22740	93.52
PGI	lavender 1821	7.33	12	95.77

TABLE 5.1: Costs, time spendings and results of SST2. †means multiprocessing (5 processes) is enabled. Time for manual labeling excludes the time spent on instruction preparation and training.

in better performance. This is because entities are not typically key factors in the sentiment classification task, as most entities are neutral and do not provide additional information relevant to sentiment. Furthermore, since a large portion of the data in SST2 does not contain any entities, the sentences generated using DADG do not follow the same distribution as the test data in SST2, leading to poorer performance. For comparison purposes, the result of PGI is also presented. It is suggested that, for small-scale applications, it is practical to use GPT-3 to directly label unlabeled data.

5.4.2.2 FewRel

The FewRel dataset is used for RE experiments. The original FewRel dataset, proposed for meta-learning, is re-formulated to a supervised learning setting. The train data of FewRel, which comprises 64 distinct relations and 700 labeled instances for each relation, is divided into a new train/dev/test split (560/70/70). It is to simulate the real-world application of GPT-3 to annotate data for tasks with large label spaces. For FewRel experiments, we follow [251] to fine-tune $\text{BERT}_{\text{BASE}}$ on the data created by the three approaches for 3 epochs. Subsequently, the fine-tuned model is evaluated on the human-labeled test data to assess the quality of data produced by the proposed approaches. The number of samples annotated or generated by each approach is determined by assuring the costs of each approach are comparable.

Annotation Approaches The FewRel dataset poses significant challenges for the PGDA approach, primarily due to the complexity of instructing GPT-3 to comprehend the 64 relations. Due to the cost and maximum token length constraints of the GPT-3 API, we can only include 1-shot data for each relation within the prompt, which can make it difficult for GPT-3 to "understand" each relation. To address these challenges, we try 5 different prompts for PGDA, with the goal of exploring whether different prompts could be effective for tasks with large label space. As mentioned in Section 5.3.2, in PGDG, we conduct the annotation for RE in two steps. The first step is to instruct GTP-3 to generate head-tail entity pairs for a specified relation and the second step is to generate sentences with the generated triplets. We generate 200 labeled data for each relation. As mentioned in Section 5.3.3, DADG for RE is also conducted in two steps. The first step is to query WikiData to obtain head-tail entity pairs for a specified relation and the second step is to generate sentences with the generated triplets. We generate 200 labeled data for each relation.

Results Table 5.2 presents the results of three different approaches. All five proposed prompts for PGDA perform badly on the FewRel task due to the task difficulty and large label space. In contrast, the generation-based approaches, namely PGDG and DADG, achieve much better performance with comparable costs. Even with access to only 1-shot data, PGDG and DADG yield F1 scores of around 44 and 40 points respectively in comparison to PGDA. With access to 5-shot data, the performances of PGDG and DADG are further improved with the increased diversity of the generated data. Under comparable costs, PGDG and DADG outperform the human-labeled data (704 data points) with 33-point and 23-point F1 scores respectively. It is worth noting that the PGDG approach consistently outperforms the DADG approach. Through analysis, it is determined that the head-tail entity pairs generated by PGDG possess greater diversity than those generated by DADG for specific relations such as religion and the language of the work. We do not perform PGI on FewRel data as the cost is obviously much higher.

Approach	Num. of Samples	Cost (USD)	Time (Mins)	P	R	F1
PGDA1 (1-shot)	384	28.55	13†	0.03	1.56	0.05
PGDA2 (1-shot)	384	25.40	10†	0.14	1.7	0.18
PGDA3 (1-shot)	384	25.19	11†	0.09	1.65	0.13
PGDA4 (1-shot)	384	25.57	10†	0.02	1.56	0.05
PGDA5 (1-shot)	384	25.56	11†	0.02	1.56	0.05
PGDG (1-shot)	12800	30.58	285†	47.82	45.58	**44.11**
DADG (1-shot)	12800	17.16	220†	45.41	42.41	40.02
PGDG (5-shot)	12800	99.35	340†	70.59	67.99	**67.71**
DADG (5-shot)	12800	88.91	265†	59.76	60.85	57.98
Human Labeled	palepink 704	101 - 200	640	41.92	41.45	34.22
	palepink 12800	1828 - 3584	11636	85.19	85.07	84.95
	palepink 35840	6400 - 10,000	32582	87.55	87.43	87.34
PGI	lavender -	-	-	-	-	-

TABLE 5.2: Costs, time spendings, and results of FewRel. Time for manual labeling excludes the time spent on instruction preparation and training. The number of samples annotated or generated by each approach is determined by assuring **comparable costs**. †means multiprocessing (5 processes) is enabled.

5.4.3 Token-Level Task

5.4.3.1 CrossNER

The AI domain split in CrossNER has 14 entity classes, namely product, field, task, researcher, university, programming language, algorithm, misc, metrics, organisation, conference, country, location, person. We fine-tune $\text{BERT}_{\text{BASE}}$ on the CrossNER task with corresponding data for 100 epochs with early stopping.

Annotation Approaches In PGDA for each entity type, we initiate GPT-3 to generate its definition and provide a selection of data (no more than 10-shot) with entities belonging to the specified entity type in the prompt to assist GPT-3 in recognizing entities belonging to the same class within the unlabeled data. It is observed that the same entity may be labeled as different entity types with different prompts. Therefore, we also include an additional prompt to determine the final entity type for each identified entity. Both PGDG and DADG for CrossNER are conducted in two steps. The first step for PGDG is to prompt GPT-3 to generate entities for each entity type. On the other hand, the first step for DADG is to query Wikidata to get the entities of each entity type. Notice that we use no more than 200 generated entities for each entity type in our experiments for both PGDG and DADG. The second step of both approaches is to use the generated entities to generate sentences within a specific domain using GPT-3. In the process of

Approach	Num. of Samples	Cost (USD)	Time (Mins)	Results
PGDA (10-shot)	100	15.39	21	23.08
PGDG (Zero-shot)	1500	7.78	17†	42.63
	3000	13.56	33†	41.35
DADG (Zero-shot)	1500	6.77	20†	46.90
	3000	13.61	40†	**47.22**
Human Labeled	palepink 100	17 - 42.85	65	42.00
PGI	lavender 431	63.23	20†	46.65

TABLE 5.3: Cost, time spending and results of CrossNER (AI Domain Split). Time for manual labeling excludes the time spent on instruction preparation and training. †means multiprocessing (5 processes) is enabled.

generating sentences for both PGDG and DADG, we randomly select a few entities from all the entities to generate each sentence.

Results Table 5.3 presents the results of the three approaches. We find the train data labeling method using PGDA has the worst performance yet the highest costs among the three proposed approaches. It should be noted that there are only 100 gold train data points in the AI domain split in the CrossNER dataset, and these same 100 data points are labeled using PGDA. However, the cost of labeling these 100 data points is higher than the cost of using the generation approaches to generate 3000 data points. It is observed that GPT-3 is effective at identifying entities in the text, but it may also identify entities that are not of the specified entity type, resulting in incorrect labeling. Additionally, GPT-3 may not accurately identify the boundaries of the entities. These two disadvantages make it impractical to use PGDA for labeling data for named entity recognition (NER) in a production setting, especially when the label space becomes bigger. The PGDG approach is able to achieve a result comparable to the 100 human-labeled gold train data at a lower cost. When utilizing Wikidata, the DADG approach is able to achieve a higher result than PGDG, likely due to its ability to leverage more unique entities and in-domain entities extracted from Wikidata. This shows that the ability to access in-domain entities is crucial for creating high-quality training data for NER.

5.4.3.2 ASTE

We follow [293] to fine-tune $\text{BERT}_{\text{BASE}}$ on the ASTE task using data created by each approach for 10 epochs and evaluate the fine-tuned models on human-labeled

Approach	Num. of Samples	Cost (USD)	Time (Mins)	P	R	F1
PGDA1	906	11.34	18	57.93	44.38	**50.26**
PGDA2	906	9.02	17	50.78	24.13	32.71
PGDA3	906	12.84	19	50.73	38.31	43.65
PGDG1	1000	9.41	15†	44.36	22.47	29.83
PGDG2	1000	7.68	14†	54.93	14.36	22.77
PGDG3	1000	13.77	18†	45.10	12.71	19.83
DADG	1000	13.74	18†	48.61	6.45	11.38
Human Labeled	palepink91	13 - 20	180	45.14	38.49	41.55
	palepink 906	130 - 200	1800	63.07	55.99	59.32
PGI	lavender328	3.92	9	50.10	48.43	49.25

TABLE 5.4: Costs, time spendings and results of ASTE (laptop domain split). Time for manual labeling excludes the time spent on instruction preparation and training. †means multiprocessing (5 processes) is enabled.

test data. We conduct our experiment under 10-shot settings.

Annotation Approaches In PGDA, we randomly sample 10-shot data from gold train data and use them to guide GPT-3 to tag the unlabeled data. Given the complexity of ASTE, which requires the identification of aspect, opinion, and sentiment triplets, we try 3 different prompts to assess the impact of different prompts on the overall performance of the tagging process. In PDGD, for comparison purposes, the same 10-shot data used for PGDA is used in the experiments for PGDG. We first instruct GPT-3 to generate aspect-opinion-sentiment triplets and then instruct GPT-3 to generate sentences with the generated triplets. We also try on 3 prompts under PGDG. In DADG, we query entities in laptop and computer hardware domains from WikiData and used them as aspects. We use the prompt that achieved the best performance for PGDG as the prompt to generate opinions and sentiments for the aspects. Then we use the obtained triplets for sentence generation.

Results Table 5.4 presents the results of three different approaches. PGDA achieves the best performance compared with the other approaches. We also notice that performance varies with different prompts, which aligns with the previous research [294]. Similar to SST2, as entities are not the key factors for ASTE and provide little help to this task, DADG is also outperformed by PGDA.

5.5 Further Analysis

5.5.1 Impact of Label Space

The results of our experiments indicate that the tagging-based approach (PGDA) is more appropriate for tasks with smaller label spaces and clearly defined labels. Examples of such tasks include sentence-level sentiment analysis and ASTE, which both have small label space (2-3 labels) that can be easily distinguished, e.g. positive, negative, neutral. In contrast, the generation-based approaches (PGDG and DADG) are better suited for tasks with larger label spaces or labels that possess a certain degree of ambiguity. Examples of such tasks include CrossNER and FewRel, which have 14 and 64[5] labels respectively, and some of which may be difficult to identify or differentiate (e.g. Misc, etc.). Both the tagging-based and generation-based approaches have their own advantages and disadvantages. The tagging-based approach allows for direct access to in-domain unlabeled data, while the generation-based approaches may generate data that contains information that was "learned" during pre-training and may not align with the distribution of in-domain data. However, as the label space becomes larger, the tagging-based approach requires a lengthy prompt with examples to guide GPT-3, which can lead to catastrophic forgetting and increase annotation costs. On the other hand, the generation-based approaches can reformulate the task by first generating spans with labels (e.g. entities and triplets), and then generating a sentence with the labeled spans. These approaches reduce label errors and avoid the challenges of span boundary detection. In addition, generation-based approaches tend to be more cost-effective. as the prompts used can be significantly shorter when compared to those used in the tagging-based approach and multiple data can be generated with a single prompt at a time.

5.5.2 Comparision with Human Annotators

Through the extensive experiments we find that GPT-3 demonstrates promising ability to generate domain-specific data (e.g., entities in AI), structured data (e.g., triplets), as well as unstructured sequences at a fast speed. As discussed above,

[5]We refer to the train split of the FewRel used in our experiments. The original FewRel data has 100 labels in total.

Generated Entities: Chiang Mai International Airport; Chiang Mai, Thailand;
Generated Sentence: Chiang Mai International Airport is the main gateway for air travels to and from Chiang Mai, Thailand.

FIGURE 5.5: An example to demonstrate the generation ability of GPT-3.

GPT-3 can even be used to generate data from scratch or to convert structured knowledge into natural sentences (Figure 5.5), eliminating the requirement of unlabeled data. While for human annotators, it usually takes longer time to train them for domain-specific data annotation, and their annotation speed is not comparable with machines in most cases. Moreover, it is often more challenging for human to construct training data without unlabeled data, or when the size of label space is very large. Therefore, in terms of speed and domain-specific data annotation, and in the setting of labeled data generation, large language models (LLMs) exhibits encouraging potential. Machines are good at quickly labeling or generating a large amount of training data. However, if we limit the number of data samples for model training, the per-instance quality of the data annotated by human is still higher in most cases.

5.5.3 Impact of Number of Shots

We conduct experiments on the following two datasets, SST2 and FewRel to explore the impact of the number of shots. We find that increasing the number of shots does not necessarily lead to better annotation results for all approaches. As shown in Figure 5.6, for SST2, tagging approach (PGDA) can benefit from more examples in the context, which enhances GPT-3's ability to tag unlabeled data. However, for the PGDG and DADG approaches, GPT-3 tends to generate data similar to the given examples. As shown in Figure 5.7, for SST2, the data is usually not a complete sentence and tend to be short and carry less information. Thus, with more data examples, GPT-3 will "learn" to generate similar data with less information and lead to poorer data quality. However, for FewRel, the data is a complete sentence and carry lots of information and the relations between the head entity and tail entity tend to be more implicit. Thus, with 5-shot data in the context,

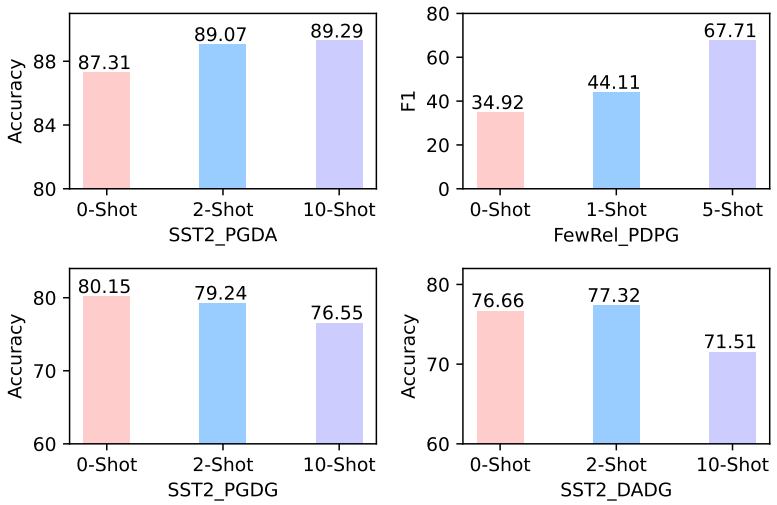


FIGURE 5.6: Experiments on the impact of number of shots. We reported the results of 6,000 data on SST2 and 12,800 data (200 data per class) on FewRel.

SST Example: a smile on your face (positive)
FewRel Example: Winscombe is a lightly populated locality in the southern part of the Canterbury region of New Zealand 's South Island . (Relation: located on terrain feature)

FIGURE 5.7: Examples to show the differences between the data distributions of SST2 and FewRel data.

GPT-3 can generate data that also contain more implicit relations than only with 1-shot or zero-shot in the context.

5.6 Chapter Summary

In this chapter, we conducted a comprehensive examination of the utility of GPT-3, a state-of-the-art large language model, as a tool for annotating data across a range of natural language processing (NLP) tasks. This investigation is anchored in three primary methodologies. Our in-depth experimental analysis reveals that GPT-3 exhibits promising capabilities in the realm of data annotation for diverse NLP tasks. This is particularly significant for entities such as individuals or organizations operating under financial constraints, as the cost associated with GPT-3's annotation services is relatively modest compared to other options.

A striking observation from our experiments is the performance level of models that are trained using data annotated by GPT-3. In many instances, these models demonstrate performance metrics that are on par with, or in some cases surpass, models trained on datasets annotated by human experts. This finding is noteworthy, especially considering the reduced financial investment required for GPT-3's services. However, it's imperative to acknowledge that while GPT-3's annotation quality is commendable, it still lags behind the precision and nuance often found in human-annotated datasets. There remains a discernible gap in quality that future enhancements and iterations of the model could potentially bridge.

The implications of our findings are far-reaching and shine a spotlight on the potential for automated data annotation using advanced language models such as GPT-3. We believe that the insights gleaned from our work will stimulate further research and the development of innovative methods aimed at improving the quality of data generated by these sophisticated models. The ultimate goal is to refine these tools to a point where they can reliably produce high-quality annotated data, thus democratizing the field of AI. By enabling a broader spectrum of individuals and organizations to generate their own data for model training, we are stepping into an era where the democratization of AI becomes a tangible reality. This shift holds the promise of leveling the playing field, allowing a wider community to participate in and benefit from the advancements in AI technology.

Chapter 6

Conclusions and Future Directions

6.1 Overall Summary

This thesis presents a comprehensive study on enhancing information extraction and dialogue system performance in low-resource scenarios. It consists of several research projects, each focusing on a unique aspect of this challenge.

Chapter 3 introduces a generation-based data augmentation technique specifically designed for low-resource sequence tagging tasks. This innovative approach uses advanced generative models to create diverse and representative datasets, addressing data scarcity issues. The chapter demonstrates how this method significantly improves the accuracy and robustness of information extraction systems, particularly in languages and domains with sparse data. This research has inspired follow up research on the token-level data augmentation and generation-based data augmentation for NLP.

In Chapter 4, the research shifts to the development of a multilingual task-oriented dialogue system in the context of globalization. This system integrates local entities and contexts within a global framework. The chapter discusses the challenges and solutions in creating systems that operate effectively across various languages and cultural contexts. It also highlights how incorporating local nuances can enhance the system's performance and relevance in a globalized environment. This research

has inspired follow up research on the multilingual and foreign langauges dialogue systems.

Chapter 5 explores the application of large language models for improved data labeling in natural language processing tasks. This chapter delves into how these models can automate and refine the data annotation process, significantly reducing the time and resources needed for manual labeling. It emphasizes the effectiveness of these models in generating high-quality, reliable labels in complex linguistic scenarios. This has inspired research of using LLMs generated data for model training and human-AI collaboration of data annotations.

6.2 Future Directions

In the current section, we aim to comprehensively delineate a plethora of prospective avenues for future research endeavors. This includes, but is not limited to, the strategic expansion of previously established methodologies, enhancing their applicability and versatility across a more expansive array of tasks and disciplines. In addition, we emphasize the necessity of addressing and resolving the substantial challenges inherent in the most recent and advanced techniques that represent the forefront of our field.

Our intent is to catalyze further scholarly inquiry by broadening the scope of existing strategies, thereby enabling them to adapt to and effectively manage a diverse range of scenarios and problem sets. This expansion is not merely a matter of scaling up current methods, but involves a thoughtful reconfiguration and refinement to ensure they are robust, efficient, and suitable for a wider spectrum of applications.

Moreover, we recognize the imperative to confront and overcome significant hurdles present in state-of-the-art methodologies. These cutting-edge techniques, while representing the zenith of current research, often harbor complexities and issues that must be meticulously examined and resolved. This entails a deep dive into their foundational principles, operational mechanisms, and potential limitations. By doing so, we can enhance their reliability, efficacy, and applicability, thereby advancing the frontier of knowledge in our field.

In summary, this section is dedicated to charting a course for future research that is both ambitious and pragmatic. It calls for a dual approach: extending the reach of existing methods while simultaneously refining the latest technological innovations to address their inherent challenges. Through this dual focus, we aim to forge a path forward that is innovative, comprehensive, and conducive to significant advancements in our field of study.

6.2.1 Multilingual Logical and Mathematical Reasoning

In the field of artificial intelligence research, a promising avenue for future exploration lies in augmenting the capabilities of AI systems to proficiently comprehend and perform logical and mathematical reasoning in a diverse array of languages [295–297]. This task goes beyond simple translation; it involves a deep understanding of the cultural and linguistic nuances that influence different ways of reasoning. At the heart of this challenge is the need to create AI models that are adept at adjusting to the varied logical structures and mathematical standards that are inherent in different language environments. This requires a nuanced approach that takes into account the unique aspects of each language and culture [276, 298, 299].

In addition, there's a significant opportunity to delve into the interplay between established frameworks of formal logic and the field of natural language processing [300]. By integrating these two areas, AI systems could achieve a higher level of precision in reasoning, especially in scenarios where multiple languages are involved. This exploration could lead to groundbreaking developments in how AI understands and processes different forms of logic and reasoning across various languages. The focus would be on enabling AI to not just translate words, but to truly grasp the underlying logical constructs and mathematical principles that are expressed differently in each language. This advancement would represent a major leap in the AI's ability to interact with and understand a multilingual world, breaking new ground in the field of artificial intelligence.

6.2.2 Culture-aware Multilingual NLP

The exploration within this particular domain may shift focus towards the creation and development of Natural Language Processing (NLP) systems that are not

only equipped to handle multiple languages simultaneously but also possess a keen sensitivity to the various cultural aspects embedded within these languages [301–303]. The primary goal of these sophisticated systems would be to go beyond the realm of simple, direct translations. They would be engineered to delve deeply into understanding and accurately interpreting the nuances of idioms, the subtle meanings behind cultural references, and the unique contextual implications that are specific to each language and its associated culture [304].

This would involve an advanced level of linguistic analysis where the systems are not just translating words but are also interpreting the cultural context in which these words are used. The essence of this endeavor is to capture the rich, cultural underpinnings that often get lost in translation, ensuring that the true intent and flavor of the original message are conveyed in the target language.

In addition to this, a significant and promising direction for research could be the formulation and implementation of algorithms that are highly skilled in identifying, understanding, and adapting to various cultural biases and nuances. This aspect of research is crucial because each language and culture comes with its own set of biases and intricacies, which can significantly impact the meaning and interpretation of text. By recognizing and adjusting for these cultural biases, these advanced NLP systems would be able to provide more accurate, contextually relevant translations and interpretations.

This would elevate the functionality and effectiveness of NLP technologies in a wide range of international and multicultural settings [305]. By enhancing their ability to accurately interpret and adapt to different cultural contexts, these systems could greatly improve communication across languages and cultures, making them invaluable tools in our increasingly globalized world. The potential applications of such technology are vast, ranging from improving international business communications to aiding in cross-cultural understanding and collaboration.

6.2.3 Domain Adaptation in Multilingual Large Language Models (LLMs)

Future research endeavors in the field of artificial intelligence and natural language processing could focus on enhancing the domain-specific flexibility of large language models [306]. This involves ensuring that these models remain effective and reliable when inferencing across a multitude of languages. The primary objective in this advancement would be to design and develop models that are adept at understanding, assimilating, and utilizing industry-specific jargon and concepts in a variety of languages. By achieving this, the need for extensive retraining of these models for each new domain or language could be significantly reduced [307].

One of the foremost challenges in this pursuit would be addressing the scarcity of data, particularly in specialized fields, for languages that are not widely spoken. This is a crucial aspect because the effectiveness of language models often hinges on the availability and richness of the data they are trained on. Another significant hurdle would be the development and refinement of advanced transfer learning techniques. These techniques are essential for enabling the models to efficiently transfer and apply the knowledge acquired from one language or domain to another, essentially learning from one context and applying that learning to a different context. This process requires sophisticated algorithms capable of identifying and leveraging the commonalities and differences between various languages and domains.

Moreover, this research could delve into the intricacies of cross-lingual and cross-domain nuances, ensuring that the models are not only proficient in multiple languages but also sensitive to the cultural and contextual subtleties inherent in each language. This would involve the models being trained to understand idiomatic expressions, colloquialisms, and cultural references, which vary greatly from one language to another. The aim is to create models that are not only linguistically versatile but also culturally informed, thereby enhancing their applicability and relevance in a global context.

The success of such research could revolutionize the way we interact with technology, making language models more accessible, efficient, and useful across a wider range of industries and linguistic communities. It could lead to breakthroughs in how language models are used for tasks such as translation, content creation,

and even in understanding and predicting human behavior and preferences across different cultures and languages.

6.2.4 Human-AI Interaction

The remarkable achievements of ChatGPT can be seen as a testament to the broader triumphs in the field of Human-AI Interaction (HAI) [308]. This particular line of research endeavors to delve deeper into enhancing the cooperative capabilities of AI in numerous activities, including but not limited to co-authoring documents, programming, data co-annotation, the art of persuasion, and even in roles akin to counselling [309]. The central objective of this research is to cultivate AI systems that possess a profoundly sophisticated grasp of human instructions and inputs [310]. Such systems would be adept at not only understanding but also anticipating human needs. They would be capable of offering creative and out-of-the-box suggestions while responding in a manner that reflects emotional intelligence.

An essential aspect of this research involves investigating how AI can modify its methods of communication and interaction to better resonate with the unique preferences and cultural nuances of individual users. This approach is critical because it recognizes the diverse nature of human users and their varying requirements. By doing so, AI can become more effective in its collaborations with humans, enhancing the overall experience and productivity in a wide array of tasks.

Further exploration in this domain could potentially lead to AI systems that are not just reactive but proactive in their interactions with humans. For instance, in a co-writing scenario, such an AI would not only assist in generating text based on direct prompts but also suggest alternative angles, introduce novel concepts, and even predict questions or concerns a human collaborator might have. In programming, the AI could foresee potential bugs or inefficiencies in code and propose optimizations or corrections even before the human programmer identifies them.

In persuasive tasks, AI could learn to understand and adapt to the psychological and emotional states of the human counterpart, thereby tailoring its arguments or suggestions in a way that is more likely to be well-received. In a counselling role, the AI could provide support and advice that is empathetic and tailored to

the individual's emotional state and personal history, potentially even recognizing subtle cues in the user's language and behavior that might indicate their mood or needs.

Key to all these advancements is the development of AI systems that are deeply attuned to the nuances of human language, culture, emotions, and thought processes. This would require extensive research into natural language processing, cultural sensitivity algorithms, emotional intelligence modeling, and predictive analytics. The ultimate goal is to create AI partners that are not only efficient but also empathetic and responsive to the full spectrum of human diversity, thereby enhancing the quality and effectiveness of human-AI collaborations in a multitude of fields.

List of Publications[1]

Peer-reviewed Papers

- Is GPT-3 a Good Data Annotator? (ACL 2023)
  Bosheng Ding*, Chengwei Qin*, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing.

- GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems. (ACL 2022)
  Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao.

- MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. (ACL 2021)
  Linlin Liu*, **Bosheng Ding***, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao.

- DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. (EMNLP 2020)
  Bosheng Ding*, Linlin Liu*, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao.

- Chain of Knowledge: A Framework for Grounding Large Language Models with Structured Knowledge Bases. (ICLR 2024)
  Xingxuan Li, Ruochen Zhao, Yew Ken Chia, **Bosheng Ding**, Lidong Bing, Shafiq Joty, Soujanya Poria

- Retrieving Multimodal Information for Augmented Generation: A Survey. (EMNLP 2023)

[1]The superscript * indicates joint first authors

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Long, Chengwei Qin, **Bosheng Ding**, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty.

- On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. (ACL 2021)
  Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, **Bosheng Ding**, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si.

Preprint Papers

- Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines
  Ruochen Zhao*, Xingxuan Li*, Yew Ken Chia*, **Bosheng Ding***, Lidong Bing.

- LogicLLM: Exploring Self-supervised Logic-enhanced Training for Large Language Models. (Under review)
  Fangkai Jiao, Zhiyang Teng, Shafiq Joty, **Bosheng Ding**, Aixin Sun, Zhengyuan Liu, Nancy F Chen.

- Panda LLM: Training Data and Evaluation for Open-Sourced Chinese Instruction-Following Large Language Models. (Technical Report)
  Fangkai Jiao*, **Bosheng Ding***, Tianze Luo*, Zhanfeng Mo*.

Bibliography

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[2] Jack Cowan. Neural networks: the early days. *Advances in neural information processing systems*, 2, 1989. 1

[3] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 1

[4] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 1

[5] Anthony TC Goh. Back-propagation neural networks for modeling complex systems. *Artificial intelligence in engineering*, 9(3):143–151, 1995. 1

[6] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020. 1

[7] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pages 128–144. Springer, 2020. 1

[8] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019. 1

[9] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56 (7):5975–6037, 2023. 1

[10] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480, 2002. 1

[11] Yunfei Lai. A comparison of traditional machine learning and deep learning in image recognition. In *Journal of Physics: Conference Series*, volume 1314, page 012148. IOP Publishing, 2019. 1

[12] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021. 1

[13] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1, 3, 13

[14] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8, 2017. 2

[15] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34, 2021. 2

[16] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.84. URL `https://aclanthology.org/2021.findings-acl.84`. 2, 3

[17] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211, 2023. 2

[18] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020. 3, 46, 47

[19] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015. 3

[20] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670. URL `https://aclanthology.org/D19-1670`. 3, 14

[21] Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulić. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2104.08570*, 2021. URL `https://arxiv.org/abs/2104.08570`. 4, 36, 70

[22] Isabelle Buchstaller. The localization of global linguistic variants. *English World-Wide*, 29(1):15–44, 2008. 4

[23] Tae Soo Kim, Geonwoon Jang, Sanghyup Lee, and Thijs Kooi. Did you get what you paid for? rethinking annotation cost of deep learning based computer aided detection in chest radiographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 261–270. Springer, 2022. 5

[24] Alex Hernández-García and Peter König. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*, 2018. 13

[25] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021. 14

[26] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015. 14

[27] Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. Data boost: Text data augmentation through reinforcement learning guided conditional generation. 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 9031–9041, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.726. URL `https://aclanthology.org/2020.emnlp-main.726`. 14

[28] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao (Bernie) Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54:1 – 40, 2020. URL `https://api.semanticscholar.org/CorpusID:221397441`. 14

[29] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H1aIuk-RW`. 15

[30] Giuseppe Riccardi and Dilek Z. Hakkani-Tür. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13:504–511, 2005. URL `https://api.semanticscholar.org/CorpusID:1635495`. 15

[31] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *International Conference on Machine Learning*, 1999. URL https://api.semanticscholar.org/CorpusID:1371723. 15

[32] William H. Beluch, Tim Genewein, A. Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018. URL https://api.semanticscholar.org/CorpusID:52838058. 15

[33] Samuel Budd, Emma Claire Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71:102062, 2019. URL https://api.semanticscholar.org/CorpusID:203837078. 15

[34] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27:2591–2600, 2017. URL https://api.semanticscholar.org/CorpusID:206663375. 15

[35] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL https://api.semanticscholar.org/CorpusID:7200347. 15

[36] Jianping Gou, B. Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129: 1789 – 1819, 2020. URL https://api.semanticscholar.org/CorpusID:219559263. 15, 16

[37] S. Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:201670719. 15

[38] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2019. URL https://api.semanticscholar.org/CorpusID:204788964. 15

[39] Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7130–7138, 2017. URL https://api.semanticscholar.org/CorpusID:206596723. 16

[40] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. *Proceedings of machine learning research*, 139:12878–12889, 2021. URL `https://api.semanticscholar.org/CorpusID:235125689`. 16

[41] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018. 16

[42] Changlin Li, Jiefeng Peng, Liuchun Yuan, Guangrun Wang, Xiaodan Liang, Liang Lin, and Xiaojun Chang. Block-wisely supervised neural architecture search with knowledge distillation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1986–1995, 2019. URL `https://api.semanticscholar.org/CorpusID:208513081`. 16

[43] Samuel Stanton, Pavel Izmailov, P. Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Neural Information Processing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:235390933`. 16

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL `https://api.semanticscholar.org/CorpusID:53592270`. 17

[45] Li Zhou, Qinwei Fan, Xiaodi Huang, and Yan Liu. Weak and strong convergence analysis of elman neural networks via weight decay regularization. *Optimization*, 72:2287 – 2309, 2022. URL `https://api.semanticscholar.org/CorpusID:247961182`. 17

[46] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997. URL `https://api.semanticscholar.org/CorpusID:9219592`. 17

[47] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. URL `https://api.semanticscholar.org/CorpusID:28671436`. 18

[48] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Neural Information Processing Systems*, 1991. URL `https://api.semanticscholar.org/CorpusID:10137788`. 18

[49] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. URL `https://api.semanticscholar.org/CorpusID:538820`. 19

[50] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks

from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. URL https://api.semanticscholar.org/CorpusID:6844431. 19

[51] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Neural Information Processing Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:15953218. 20

[52] Jitendra Kumar, Ashutosh Kumar Singh, Anand Mohan, and Rajkumar Buyya. Ensemble learning. *Machine Learning Foundations*, 2020. URL https://api.semanticscholar.org/CorpusID:9963037. 20

[53] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, 2018. URL https://api.semanticscholar.org/CorpusID:49291826. 20

[54] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241 – 258, 2019. URL https://api.semanticscholar.org/CorpusID:201667785. 21

[55] Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 2000. URL https://api.semanticscholar.org/CorpusID:56776745. 21

[56] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010. URL https://api.semanticscholar.org/CorpusID:740063. 22

[57] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL https://api.semanticscholar.org/CorpusID:204838007. 22, 104

[58] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76, 2019. URL https://api.semanticscholar.org/CorpusID:207847753. 22

[59] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, 2018. URL https://api.semanticscholar.org/CorpusID:51929263. 22

[60] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online,

July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747. 23, 72

[61] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, Mar 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00288. URL http://dx.doi.org/10.1162/tacl_a_00288. 23

[62] Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:218470133. 23

[63] Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. The sigmorphon 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. *ArXiv*, abs/1910.11493, 2019. URL https://api.semanticscholar.org/CorpusID:201679015. 23

[64] Sebastian Schuster, S. Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *North American Chapter of the Association for Computational Linguistics*, 2018. URL https://api.semanticscholar.org/CorpusID:53110354. 24

[65] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Conference on Empirical Methods in Natural Language Processing*, 2017. URL https://api.semanticscholar.org/CorpusID:9489563. 24

[66] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *North American Chapter of the Association for Computational Linguistics*, 2018. URL https://api.semanticscholar.org/CorpusID:67856712. 24

[67] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. Interactive attention transfer network for cross-domain sentiment classification. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:13822774. 25

[68] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760, 2010. 26

[69] Yang Zou, Xiaodong Yang, Zhiding Yu, B. V. K. Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *European Conference on Computer Vision*, 2020. URL https://api.semanticscholar.org/CorpusID:220646469. 26

[70] Wouter M. Kouw and M. Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:766–785, 2019. URL https://api.semanticscholar.org/CorpusID:198898096. 26

[71] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *ArXiv*, abs/1610.05755, 2016. URL https://api.semanticscholar.org/CorpusID:8696462. 26

[72] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *ArXiv*, abs/2007.02454, 2020. URL https://api.semanticscholar.org/CorpusID:220363892. 26

[73] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:8239952. 26

[74] Richard Socher, Milind Ganjoo, Christopher D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Neural Information Processing Systems*, 2013. URL https://api.semanticscholar.org/CorpusID:2808203. 27

[75] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2251–2265, 2017. URL https://api.semanticscholar.org/CorpusID:4852047. 27

[76] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 2015. URL https://api.semanticscholar.org/CorpusID:5891792. 27

[77] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. URL https://api.semanticscholar.org/CorpusID:232035663. 27

[78] Yingke Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011. URL https://api.semanticscholar.org/CorpusID:196536. 28

[79] Oscar Day and Taghi M. Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4, 2017. URL `https://api.semanticscholar.org/CorpusID:10092092`. 28

[80] Qiang Yang, Yuqiang Chen, Gui-Rong Xue, Wenyuan Dai, and Yong Yu. Heterogeneous transfer learning for image clustering via the socialweb. In *Annual Meeting of the Association for Computational Linguistics*, 2009. URL `https://api.semanticscholar.org/CorpusID:16197334`. 28

[81] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2017. URL `https://api.semanticscholar.org/CorpusID:4412459`. 28

[82] Yaqing Wang, Quanming Yao, James Tin-Yau Kwok, and Lionel Ming shuan Ni. Generalizing from a few examples: A survey on few-shot learning. *arXiv: Learning*, 2019. URL `https://api.semanticscholar.org/CorpusID:226931458`. 28

[83] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Neural Information Processing Systems*, 2018. URL `https://api.semanticscholar.org/CorpusID:44061218`. 28

[84] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 403–412, 2018. URL `https://api.semanticscholar.org/CorpusID:54448258`. 28

[85] Han-Jia Ye, Hexiang Hu, De chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8805–8814, 2018. URL `https://api.semanticscholar.org/CorpusID:214713930`. 29

[86] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8239–8247, 2018. URL `https://api.semanticscholar.org/CorpusID:54459283`. 29

[87] Isabelle M Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003. URL `https://api.semanticscholar.org/CorpusID:379259`. 29

[88] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. *ArXiv*, abs/2010.07877, 2020. URL `https://api.semanticscholar.org/CorpusID:222378404`. 29

[89] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19325–19337, 2023. URL `https://api.semanticscholar.org/CorpusID:255942320`. 30

[90] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guan Wang, Kaichao Zhang, Cheng Ji, Qi Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, Lichao Sun Michigan State University, Beihang University, Lehigh University, Macquarie University, Nanyang Technological University, University of California at San Diego, Duke University, University of Chicago, and Salesforce AI Research. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *ArXiv*, abs/2302.09419, 2023. URL `https://api.semanticscholar.org/CorpusID:257039063`. 30

[91] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018. 30

[92] Ivan Vulic, E. Ponti, Robert Litschko, Goran Glavas, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL `https://api.semanticscholar.org/CorpusID:222290596`. 30

[93] 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 (NAACL), 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`. 30, 72, 77

[94] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 31, 102

[95] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://aclanthology.org/2020.acl-main.703`. 31

[96] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 31, 32, 102, 103

[97] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. A survey of large language models. *ArXiv*, abs/2303.18223, 2023. URL `https://api.semanticscholar.org/CorpusID:257900969`. 31

[98] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022. URL `https://api.semanticscholar.org/CorpusID:249674500`. 32

[99] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ArXiv*, abs/2304.13712, 2023. URL `https://api.semanticscholar.org/CorpusID:258331833`. 32

[100] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016. 32, 48, 54, 56

[101] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0306050, 2003. 32

[102] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *North American Chapter of the Association for Computational Linguistics*, 2003. URL `https://api.semanticscholar.org/CorpusID:14835360`. 33

[103] Helmut Schmidt. Probabilistic part-of-speech tagging using decision trees. 1994. URL `https://api.semanticscholar.org/CorpusID:17392458`. 33

[104] Bernd Bohnet, Ryan McDonald, Goncalo Simoes, Daniel Andor, Emily Pitler, and Joshua Maynez. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. *arXiv preprint arXiv:1805.08237*, 2018. 33, 48

[105] Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, 2018. 33, 48

[106] Bing Liu. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012. 34

[107] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35:11019–11038, 2022. URL `https://api.semanticscholar.org/CorpusID:247218352`. 34, 35

[108] Ali Muttaleb Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. Machine learning-based sentiment analysis for twitter accounts. *Mathematical & Computational Applications*, 23:11, 2018. URL `https://api.semanticscholar.org/CorpusID:52042337`. 34

[109] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307, 2011. URL `https://api.semanticscholar.org/CorpusID:3181362`. 34

[110] Pulung Hendro Prastyo, Igi Ardiyanto, and Risanuri Hidayat. A review of feature selection techniques in sentiment analysis using filter, wrapper, or hybrid methods. *2020 6th International Conference on Science and Technology (ICST)*, 1:1–6, 2020. URL `https://api.semanticscholar.org/CorpusID:247523835`. 34

[111] 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 pretraining approach. *ArXiv*, abs/1907.11692, 2019. 34, 102

[112] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 34

[113] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*, 2013. 34, 108

[114] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018. URL `https://api.semanticscholar.org/CorpusID:5034059`. 34

[115] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`. 35

[116] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2004. URL `https://www.aclweb.org/anthology/S14-2004`. 35, 59

[117] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1002. URL `https://www.aclweb.org/anthology/S16-1002`. 35, 59

[118] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation*, 2015. URL `https://api.semanticscholar.org/CorpusID:61874237`. 35

[119] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *Conference on Empirical Methods in Natural Language Processing*, 2016. URL `https://api.semanticscholar.org/CorpusID:18993998`. 35

[120] Asha S. Manek, P. Deepa Shenoy, M. Chandra Mohan, and K. R. Venugopal. Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. *World Wide Web*, 20:135–154, 2017. URL `https://api.semanticscholar.org/CorpusID:5516330`. 35

[121] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *International Conference on Language Resources and Evaluation*, 2006. URL `https://api.semanticscholar.org/CorpusID:6247656`. 35

[122] Lu Xu, Hao Li, Wei Lu, and Lidong Bing. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 2339–2349, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.183. URL https://aclanthology.org/2020.emnlp-main.183. 35, 108

[123] Lina Maria Rojas-Barahona, Milica Gaić, Nikola Mrksic, Pei hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. A network-based end-to-end trainable task-oriented dialogue system. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2016. URL https://api.semanticscholar.org/CorpusID:10565222. 36

[124] Paweł 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 (EMNLP)*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547. 36, 37, 72, 73, 76

[125] Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. Neural belief tracker: Data-driven dialogue state tracking. In *Annual Meeting of the Association for Computational Linguistics*, 2016. URL https://api.semanticscholar.org/CorpusID:437687. 36, 37

[126] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e946209592563be0f01c844ab2170f0c-Paper.pdf. 36

[127] Jiaying Hu, Yan Yang, Chencai Chen, Liang He, and Zhou Yu. SAS: Dialogue state tracking via slot attention and slot information sharing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6366–6375, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.567. URL https://aclanthology.org/2020.acl-main.567. 36

[128] Xiujun Li, Yun-Nung (Vivian) Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. In *International Joint Conference on Natural Language Processing*, 2017. URL https://api.semanticscholar.org/CorpusID:18750779. 37

[129] Yun-Nung (Vivian) Chen, Dilek Z. Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. End-to-end memory networks with knowledge carryover

for multi-turn spoken language understanding. In *Interspeech*, 2016. URL https://api.semanticscholar.org/CorpusID:539059. 38

[130] Mike D. Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Annual Meeting of the Association for Computational Linguistics*, 2009. URL https://api.semanticscholar.org/CorpusID:10910955. 38

[131] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *ArXiv*, abs/1601.00770, 2016. URL https://api.semanticscholar.org/CorpusID:2476229. 38

[132] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Annual Meeting of the Association for Computational Linguistics*, 2016. URL https://api.semanticscholar.org/CorpusID:397533. 38

[133] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Conference on Empirical Methods in Natural Language Processing*, 2015. URL https://api.semanticscholar.org/CorpusID:2778800. 38

[134] eXascale Infolab. Relation extraction using distant supervision: a survey. 2019. URL https://api.semanticscholar.org/CorpusID:222117619. 39

[135] Zara Nasar, S. Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction. *ACM Computing Surveys (CSUR)*, 54:1 – 39, 2021. URL https://api.semanticscholar.org/CorpusID:233353895. 39

[136] Sung-Pil Choi, Seungwoo Lee, Hanmin Jung, and Sa kwang Song. An intensive case study on kernel-based relation extraction. *Multimedia Tools and Applications*, 71:741–767, 2014. URL https://api.semanticscholar.org/CorpusID:17817963. 39

[137] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR, 2021. 40

[138] Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. Controllable dialogue simulation with in-context learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.318. URL https://aclanthology.org/2022.findings-emnlp.318. 40

[139] Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. A unified dialogue user simulator for few-shot data augmentation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.277. URL `https://aclanthology.org/2022.findings-emnlp.277`. 40

[140] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. AugESC: Dialogue augmentation with large language models for emotional support conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.99. URL `https://aclanthology.org/2023.findings-acl.99`. 40

[141] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*, 2023. 40

[142] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach, 2023. 40

[143] Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. Selective in-context data augmentation for intent detection using pointwise V-information. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.107. URL `https://aclanthology.org/2023.eacl-main.107`. 41

[144] Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4convai-1.5. URL `https://aclanthology.org/2022.nlp4convai-1.5`. 41

[145] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Data augmentation for information retrieval using large language models, 2022. 41

[146] Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. Logicot: Logical chain-of-thought instruction-tuning data collection with gpt-4. *arXiv preprint arXiv:2305.12147*, 2023. 41

[147] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022. 41

[148] KaShun Shum, Shizhe Diao, and Tong Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*, 2023. 41

[149] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 41

[150] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*, 2023. 41

[151] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023. 41

[152] Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023. 42

[153] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can chatgpt reproduce human-generated labels? a study of social computing tasks, 2023. 42

[154] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 2023. 42

[155] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), July 2023. ISSN 1091-6490. doi: 10.1073/pnas.2305016120. URL `http://dx.doi.org/10.1073/pnas.2305016120`. 42

[156] Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations, 2023. 42

[157] Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint arXiv:2307.06090*, 2023. 42

[158] Parikshit Bansal and Amit Sharma. Large language models as annotators: Enhancing generalization of nlp models at minimal cost, 2023. 42

[159] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*, 2023. 42

[160] Simon Meoni, Eric De la Clergerie, and Theo Ryffel. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, 2023. 42

[161] Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. Q: How to specialize large vision-language models to data-scarce vqa tasks? a: Self-train on unlabeled images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15005–15015, 2023. 42

[162] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. *arXiv preprint arXiv:2305.03453*, 2023. 42

[163] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. DISCO: Distilling counterfactuals with large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.302. URL `https://aclanthology.org/2023.acl-long.302`. 43

[164] Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. CORE: A retrieve-then-edit framework for counterfactual data generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.216. URL `https://aclanthology.org/2022.findings-emnlp.216`. 43

[165] Damir Korenčić, Ivan Grubišić, Gretel Liz De La Peña Sarracén, Alejandro Hector Toselli, Berta Chulvi, and Paolo Rosso. Tackling covid-19 conspiracies on twitter using bert ensembles, gpt-3 augmentation, and graph nns. 2022. 43

[166] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. 43

[167] Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. When and how to paraphrase for named entity recognition? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.390. URL `https://aclanthology.org/2023.acl-long.390`. 43

[168] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023. 43

[169] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.192. URL `https://aclanthology.org/2021.findings-emnlp.192`. 43

[170] Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. *arXiv e-prints*, pages arXiv–2305, 2023. 43

[171] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 44

[172] Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media, 2023. 44

[173] Bo-Ru Lu, Nikita Haduong, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Noah A. Smith, and Mari Ostendorf. Dialgen: Collaborative human-lm generated dialogues for improved understanding of human-human conversations, 2023. 44

[174] Kechi Zhang, Ge Li, Jia Li, Zhuo Li, and Zhi Jin. Toolcoder: Teach code generation models to use apis with search tools. *arXiv preprint arXiv:2305.04032*, 2023. 44

[175] Patrice Y. Simard, Yann A. LeCun, John S. Denker, and Bernard Victorri. *Transformation Invariance in Pattern Recognition — Tangent Distance and*

Tangent Propagation, pages 239–274. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-540-49430-0. doi: 10.1007/3-540-49430-8_13. URL https://doi.org/10.1007/3-540-49430-8_13. 45

[176] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3688–3692. Ieee, 2016. 45

[177] Antonio D'Innocente, Fabio Maria Carlucci, Mirco Colosi, and Barbara Caputo. Bridging between computer and robot vision through data augmentation: a case study on object recognition. In *International Conference on Computer Vision Systems*, pages 384–393. Springer, 2017.

[178] Xiang Wang, Kai Wang, and Shiguo Lian. A survey on face data augmentation. *arXiv preprint arXiv:1904.11685*, 2019. 45

[179] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126, 2015. 45

[180] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017. 45

[181] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009. 46

[182] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017.

[183] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1091. URL https://www.aclweb.org/anthology/D17-1091.

[184] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B14TlG-RW. 46

[185] Dingquan Wang and Jason Eisner. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505, 2016. 46

[186] Iulian Vlad Serban, Alberto Garcia-Duran, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, 2016. 46

[187] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 46

[188] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*, 2018. 46

[189] Jason W Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019. 55

[190] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2020. 47

[191] Guillaume Raille, Sandra Djambazovska, and Claudiu Musat. Fast cross-domain data augmentation through neural sentence editing. *arXiv preprint arXiv:2003.10254*, 2020. 46, 47

[192] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837, 2018. 46

[193] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. Technical report, JOHNS HOPKINS UNIV BALTIMORE MD DEPT OF COMPUTER SCIENCE, 2001.

[194] Joel Mathew, Shobeir Fakhraei, and José Luis Ambite. Biomedical named entity recognition via reference-set augmented bootstrapping. *arXiv preprint arXiv:1906.00282*, 2019. 46

[195] M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. UXLA: A Robust Unsupervised Data Augmentation Framework for Cross-Lingual NLP. In *Proceedings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (ACL-IJCNLP 2021), Online, 2021. Association for Computational Linguistics. 46

[196] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 46

[197] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012. 47, 50

[198] Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1999. 48

[199] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[200] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2020. 48

[201] Hinrich Schütze. Part-of-speech induction from scratch. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 251–258. Association for Computational Linguistics, 1993. 48

[202] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1168. URL `https://www.aclweb.org/anthology/D15-1168`. 48, 49

[203] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018. 48

[204] Benjamin Heinzerling and Michael Strube. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. *arXiv preprint arXiv:1906.01569*, 2019. 48

[205] Barbara Plank, Anders Søgaard, and Yoav Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2067. URL `https://www.aclweb.org/anthology/P16-2067`. 48

[206] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of EMNLP*, pages 463–472, 2017. 49

[207] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4194–4200, 2018. URL `https://www.ijcai.org/proceedings/2018/0583.pdf`.

[208] Xin Li, Lidong Bing, Piji Li, and Wai Lam. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721, 2019. 49, 59

[209] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP*, pages 34–41, 2019. 49

[210] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019. 53

[211] Canasai Kruengkrai. Better exploiting latent variables in text modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5527–5532, 2019. 54

[212] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019. 54

[213] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 54

[214] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 54

[215] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. URL `https://www.aclweb.org/anthology/W02-2024`. 55

[216] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at*

HLT-NAACL 2003, pages 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419. 55, 60

[217] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2082. URL https://www.aclweb.org/anthology/S15-2082. 59

[218] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010. 60

[219] Narendra Kumar Gupta, Gökhan Tür, Dilek Z. Hakkani-Tür, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:213–222, 2006. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.3214&rep=rep1&type=pdf. 70

[220] Dan Bohus and Alexander I. Rudnicky. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23: 332–361, 2009. URL https://www.cs.brandeis.edu/~cs115/CS115_docs/Ravenclaw.pdf. 70

[221] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506. 70, 72, 73

[222] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 808–819, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1078. URL https://aclanthology.org/P19-1078.

[223] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.273. URL https://aclanthology.org/2020.emnlp-main.273. 72, 77

[224] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, pages 1–17, 2020. URL `https://arxiv.org/abs/2003.07490`. 70

[225] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics (TACL)*, 8: 281–295, 2020. URL `https://aclanthology.org/2020.tacl-1.19/`. 71, 73

[226] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*, 2021. URL `https://arxiv.org/abs/2106.02787`. 71, 73

[227] Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. (almost) zero-shot cross-lingual spoken language understanding. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038, 2018. URL `http://shyamupa.com/assets/pdf/papers/UFTHH18.pdf`. 71

[228] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1380. URL `https://aclanthology.org/N19-1380`.

[229] Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2479–2497, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.197. URL `https://aclanthology.org/2021.naacl-main.197`.

[230] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, pages 2950–2962, Online, April 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.eacl-main.257`. 71

[231] Li-Rong Cheng and Katharine Butler. Code-switching: a natural phenomenon vs language 'deficiency'. *World Englishes*, 8(3):293–309, 1989. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-971X.1989.tb00670.x. 71

[232] Eunhee Kim. Reasons and motivations for code-mixing and code-switching. *Issues in EFL*, 4(1):43–61, 2006. 71

[233] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020. 72

[234] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41. 72

[235] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990. URL https://aclanthology.org/H90-1021/. 73

[236] Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272, 2014. URL https://aclanthology.org/W14-4337/. 73

[237] Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5526. URL https://aclanthology.org/W17-5526. 73

[238] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1042. 73

[239] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-based user simulation for bootstrapping a pomdp dialogue

system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); Companion Volume, Short Papers*, pages 149–152, 2007. URL `https://aclanthology.org/N07-2038/`. 73

[240] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 8689–8696, 2020. URL `https://arxiv.org/abs/1909.05855`. 73

[241] Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.67. URL `https://aclanthology.org/2020.emnlp-main.67`. 73

[242] Bernd Kiefer, Anna Welker, and Christophe Biwer. Vonda: A framework for ontology-based dialogue management. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 93–105. Springer, 2021. URL `https://arxiv.org/abs/1910.00340`. 73

[243] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.13. URL `https://aclanthology.org/2020.nlp4convai-1.13`. 73

[244] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 422–428, Marseille, France, May 2020. European Language Resources Association. URL `https://www.aclweb.org/anthology/2020.lrec-1.53`. 73

[245] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 74, 104

[246] Yan Zeng and Jian-Yun Nie. Jointly optimizing state operation prediction and value generation for dialogue state tracking. *arXiv preprint arXiv:2010.14061*, 2020. URL `https://arxiv.org/abs/2010.14061`. 77

[247] Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 35–44, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.sigdial-1.4`. 77

[248] Colin Shunryu Garvey. A framework for evaluating barriers to the democratization of artificial intelligence. In *AAAI Conference on Artificial Intelligence*, 2018. 102

[249] Giovanni Rubeis, Keerthi Dubbala, and Ingrid Metzler. "democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Frontiers in Genetics*, 13, 2022. 102

[250] Andreas Bunte, Frank Richter, and Rosanna Diovisalvi. Why it is hard to find ai in smes: A survey from the practice and how to promote it. In *ICAART*, 2021. 102

[251] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 102, 104, 110

[252] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*, 2019. 102, 104

[253] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 102

[254] Su Young Kim, Hyeon ju Park, Kyuyong Shin, and KyungHyun Kim. Ask me what you need: Product retrieval using knowledge from gpt-3. *ArXiv*, abs/2207.02516, 2022. 102

[255] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli,

N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021. 103

[256] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv*, abs/2211.09085, 2022.

[257] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

[258] Sid Black, Stella Rose Biderman, Eric Hallahan, Quentin G. Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Martin Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Benqi Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. *ArXiv*, abs/2204.06745, 2022.

[259] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022.

[260] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana

Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. 103

[261] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239, 2022. 103

[262] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. 103

[263] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *ArXiv*, abs/2104.08691, 2021. 103, 104

[264] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

[265] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*, 2021. 104

[266] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022.

[267] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari,

and Vedant Misra. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858, 2022. 103

[268] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 103

[269] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. 2022. 103

[270] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. *ArXiv*, abs/2111.01998, 2021. 104

[271] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv*, abs/2110.07602, 2021. 104

[272] Teven Le Scao and Alexander M. Rush. How many data points is a prompt worth? In *North American Chapter of the Association for Computational Linguistics*, 2021. 104

[273] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723, 2021. 104

[274] Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juan-Zi Li, and Hong-Gee Kim. Prompt-learning for fine-grained entity typing. *ArXiv*, abs/2108.10604, 2021. 104

[275] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics*, 2018. 104

[276] Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq R. Joty, and Luo Si. Enhancing multilingual language model with massive multilingual knowledge triples. 2021. 104, 121

[277] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning*, 2011. 104

[278] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. *ArXiv*, abs/1704.05742, 2017. 104

[279] Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv: Machine Learning*, 2016. 104

[280] Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). In *AAAI Conference on Artificial Intelligence*, 2019. 104

[281] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *ArXiv*, abs/2106.03164, 2021.

[282] Chengwei Qin and Shafiq Joty. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.198. URL `https://aclanthology.org/2022.acl-long.198`. 104

[283] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. An empirical survey of data augmentation for limited data learning in nlp. *ArXiv*, abs/2106.07499, 2021. 104

[284] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A survey of data augmentation approaches for nlp. *ArXiv*, abs/2105.03075, 2021. 104

[285] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 104

[286] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. Daga: Data augmentation with a generation approach for low-resource tagging tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2020. 104

[287] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 104

[288] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2019. 104

[289] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 105

[290] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 105

[291] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1514. URL `https://aclanthology.org/D18-1514`. 108

[292] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *AAAI Conference on Artificial Intelligence*, 2020. 108

[293] Lu Xu, Yew Ken Chia, and Lidong Bing. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.367. URL `https://aclanthology.org/2021.acl-long.367`. 113

[294] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 2022. 114

[295] E. Ponti, Goran Glavavs, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL `https://api.semanticscholar.org/CorpusID:218470125`. 121

[296] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgJtT4tvB`.

[297] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL `https://api.semanticscholar.org/CorpusID:239998651`. 121

[298] Zhichao Duan, Xiuxing Li, Zhengyan Zhang, Zhenyu Li, Ning Liu, and Jianyong Wang. Bridging the language gap: Knowledge injected multilingual question answering. *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 339–346, 2021. URL `https://api.semanticscholar.org/CorpusID:245935213`. 121

[299] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Tim Baldwin. Cmmlu: Measuring massive multitask

language understanding in chinese. *ArXiv*, abs/2306.09212, 2023. URL https://api.semanticscholar.org/CorpusID:259164635. 121

[300] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022. URL https://arxiv.org/abs/2209.00840. 121

[301] Jing Huang and Diyi Yang. Culturally aware natural language inference. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:266176303. 122

[302] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural nlp. *ArXiv*, abs/2203.10020, 2022. URL https://api.semanticscholar.org/CorpusID:247594499.

[303] Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2023. URL https://api.semanticscholar.org/CorpusID:259342568. 122

[304] Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *ArXiv*, abs/2309.08591, 2023. URL https://api.semanticscholar.org/CorpusID:261875650. 122

[305] Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. Empowering llm-based machine translation with cultural awareness. *ArXiv*, abs/2305.14328, 2023. URL https://api.semanticscholar.org/CorpusID:258841694. 122

[306] Xujiang Liang, Zhaoquan Gu, Yushun Xie, Le Wang, and Zhihong Tian. Museda: Multilingual unsupervised and supervised embedding for domain adaption. *Knowl. Based Syst.*, 273:110560, 2023. URL https://api.semanticscholar.org/CorpusID:258498391. 123

[307] Fangkai Jiao, Bosheng Ding, Tianze Luo, and Zhanfeng Mo. Panda llm: Training data and evaluation for open-sourced chinese instruction-following large language models. *arXiv preprint arXiv:2305.03025*, 2023. 123

[308] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen Quinn, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz.

Guidelines for human-ai interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. URL `https://api.semanticscholar.org/CorpusID:86866942`. 124

[309] Qian Yang, Aaron Steinfeld, Carolyn Penstein Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. URL `https://api.semanticscholar.org/CorpusID:218483124`. 124

[310] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Comput. Hum. Behav.*, 122:106855, 2021. URL `https://api.semanticscholar.org/CorpusID:235559640`. 124