Understanding non-native Speech using SLU systems

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

SNEHAL RANJAN
2020121003
snehal.ranjan@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Understanding non-native Speech using SLU systems**" by **Snehal Ranjan**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Chiranjeevi Yarra

Acknowledgments

I would like to express my heartfelt gratitude to my advisor, Dr. Chiranjeevi Yarra for his steadfast guidance, invaluable support, and insightful mentorship throughout the entire journey of crafting this thesis. Dr. Yarra's expertise, encouragement, and patience have been instrumental in shaping my research endeavors and navigating the complexities of academic scholarship.

In addition, I extend my deepest appreciation to my parents for their unwavering support and boundless encouragement throughout the entirety of this endeavor. Their belief in my abilities, endless sacrifices, and unconditional love have been the cornerstone of my academic journey.

Furthermore, I am deeply indebted to my friends Aditya and Ayushman for their companionship and support throughout this academic quest. Their willingness to lend a listening ear, offer invaluable advice, and provide much-needed distractions during moments of academic stagnation have truly been indispensable. Their presence has transformed a solitary pursuit into a joyous and communal experience, bringing laughter and a sense of solidarity. I am immensely grateful for their friendship, which has not only lightened the burdens of scholarly pursuits but has also enriched my personal growth and imbued this journey with memorable moments and shared experiences.

I am also deeply grateful to VJS, Hari, Prince, and Aman for their enduring support and encouragement throughout this journey. Their consistent motivation, timely reminders, and willingness to provide valuable distractions during moments of stagnation have been invaluable. Their presence has been a source of inspiration, pushing me to keep moving forward and stay focused on my goals. I am truly fortunate to have such supportive individuals in my life, whose contributions have played a significant role in ensuring the progress and completion of this work.

Abstract

Spoken language understanding (SLU) systems are a critical component of modern dialog systems, enabling natural language interactions between humans and machines. However, their performance often degrades when dealing with non-native accents and grammatical errors, posing a significant barrier to widespread adoption and accessibility. This thesis presents a comprehensive investigation into the underlying causes of this performance limitation and proposes novel strategies to enhance the robustness and generalization capabilities of SLU models.

We construct SLU pipelines incorporating state-of-the-art automatic speech recognition (ASR) and natural language understanding (NLU) models, and benchmark their performance on standard datasets (ATIS, SNIPS) as well as synthetically generated data with controlled variations in accent and grammatical errors. Our empirical evaluation reveals significant performance degradation when models encounter non-native speech, highlighting vulnerabilities in capturing long-range dependencies and salient regions.

To corroborate these findings, we employ attention-based architectures and conduct targeted experiments to isolate the impact of accented speech and ungrammatical utterances. The motivation for utilizing attention-based models stems from their ability to provide interpretable insights into the model's decision-making process. By visualizing the attention weights, we can gain a better understanding of the regions in the input sequence that the model focuses on when making predictions. This interpretability is particularly valuable for identifying potential biases or vulnerabilities in how the model processes accented or ungrammatical speech. Leveraging these insights from the attention mechanism, we propose a novel data augmentation strategy that systematically introduces accent and grammatical variations during training, thereby improving the models' ability to handle such challenges. The attention visualizations guide the data augmentation process, allowing us to strategically perturb the input sequences in a way that mitigates the model's weaknesses and enhances its robustness.

Contents

Chapter Page

1 Introduction 1

2 Related Work 4

3 Effect of non-native accents on the performance of SLU Systems 7
3.1 Introduction 7
3.2 Methods 8
3.2.1 NLU models 8
3.2.2 Accented Speech data 9
3.2.3 POS Analysis 9
3.3 Results 10
3.3.1 Model Performance Evaluation 10
3.3.2 Effect of Transcription Errors on NLU Performance 11
3.3.3 POS Analysis 11
3.4 Conclusion 15

4 Effect of Grammatical Errors on the performance of SLU Systems 16
4.1 Introduction 16
4.2 Methods 17
4.2.1 Synthesizing Grammatical Errors 19
4.2.2 Intent Detection and Slot Filling Performance 19
4.2.3 POS Analysis 20
4.2.4 Attention Analysis 20
4.3 Results 22
4.3.1 Intent Detection and Slot Filling 22
4.3.2 POS Analysis 23
4.3.3 Attention Analysis 23
4.4 Conclusion 24

5 POS based augmentation for L2 learner based Errors 26
5.1 Introduction 26
5.2 Methods 27
5.3 Results 27
5.4 Conclusion 28

6 Conclusion and Future Works 29

Bibliography 32

List of Figures

Figure Page

3.1 Cascaded SLU Design 9
3.2 Accent Experiments Setup 10
3.3 Performance on ATIS with Cumulative Transcription Errors 12
3.4 Performance on SNIPS with Cumulative Transcription Errors 13
3.5 Per POS increase in Intent Error Rates 14

4.1 Workflow for the conducted analysis 18
4.2 How attention mechanism weights information from different tokens 21

5.1 Augmentation Pipeline 26

List of Tables

Table Page

3.1 NLU model performance on ATIS 11
3.2 NLU model performance on SNIPS 11

4.1 Dataset Statistics 17
4.2 NLU model performance on ATIS and SNIPS with and without grammatical errors . . 22
4.3 Cumulative Intent Accuracy and Slot F1-Score with Increased Errors on ATIS and SNIPS 23
4.4 Increase in Error Rates per POS when tested on Grammatically Erroneous data for ATIS and SNIPS 24
4.5 Mean Attention Weights per POS on ATIS and SNIPS 25

5.1 Model performance on ATIS on erroneous version of test split 28
5.2 Model performance on SNIPS on erroneous version of test split 28
5.3 Model performance with and without augmentation during training on ATIS 28
5.4 Model performance with and without augmentation during training on ATIS 28

Chapter 1

Introduction

Spoken Language Understanding (SLU) systems have gained significant traction in recent years due to their ability to facilitate natural human-machine communication. These systems play a pivotal role in accurately interpreting and processing spoken language input, making them an essential component in various applications such as digital assistants, voice-enabled interfaces, and conversational AI systems. However, a prominent challenge arises when SLU systems encounter non-native speech containing grammatical errors, accents, and patterns that deviate from the data distributions used during the training process.

Traditionally, SLU systems have been designed and optimized to process and comprehend native speech, which is generated by individuals who are native speakers of a specific language. The proficiency of these systems in understanding native speech can be attributed to the employment of extensive large-scale datasets encompassing a diverse range of native speech examples during the training phase. Furthermore, the constituent modules within SLU systems, such as the Natural Language Understanding (NLU) module, are equipped with abundant instances of native speech, enabling the statistical models to effectively learn and capture the inherent patterns, structures, and nuances of native language usage.

However, when non-native speech characterized by grammatical errors, pronunciation variations, and deviations from standard language norms is introduced, the performance of SLU systems tends to degrade significantly. Non-native speakers, influenced by the patterns and structures of their native languages, often make errors in pronunciation, grammar, or syntax, which can pose a substantial challenge for SLU systems. These errors introduce variations and complexities that the statistical models employed by SLU systems are not adequately equipped to handle, as they have been primarily trained on native speech data. Consequently, non-native grammatical errors can lead to a cascade of issues, including incorrect interpretations of the spoken language, inaccurate transcriptions, and misinterpretations of the speaker's intended meaning or intent. This can result in suboptimal performance or even complete failure of the SLU system, undermining its effectiveness and reliability in real-world scenarios involving non-native speakers.

The difficulties posed by non-native speech with grammatical errors in SLU systems stem from several underlying factors. Firstly, non-native speakers often exhibit distinct speech patterns and pronunciation

styles influenced by their native languages. These variations can manifest in various forms, such as altered vowel sounds, consonant substitutions, stress patterns, and intonation. SLU systems, which are primarily trained on native speech data, may struggle to adapt to these deviations, leading to inaccurate transcriptions and subsequent misinterpretations. Additionally, non-native speakers may inadvertently introduce grammatical errors into their speech, such as incorrect word order, verb tense agreement, or improper use of prepositions and articles. These errors can significantly alter the meaning and intent of the spoken utterance, posing a significant challenge for SLU systems trained on well-formed, grammatically correct speech samples. Furthermore, the diversity of non-native speech patterns and accents presents a formidable obstacle. Non-native speakers originate from different linguistic backgrounds, each with its unique set of phonological rules, syntactic structures, and cultural influences. This diversity introduces a wide range of variations that SLU systems must contend with, making it challenging to develop a universal solution capable of handling all non-native speech patterns effectively.

Previous research efforts in the domain of SLU have primarily concentrated on enhancing model performance by proposing novel architectures, refining existing models, or addressing specific issues related to individual modules. While these contributions have undoubtedly advanced the state of SLU technology, they do not extensively address the intricacies and unique challenges posed by non-native language inputs, particularly those characterized by grammatical errors and accents. One notable limitation of existing approaches is the heavy reliance on native speech data during the training phase. While some research has explored data augmentation techniques, such as introducing artificial noise or simulating acoustic variations, these methods do not explicitly target the unique characteristics of non-native speech, such as grammatical errors and accent-induced variations. Another limitation is the lack of comprehensive evaluation frameworks and benchmark datasets specifically designed to assess the performance of SLU systems on non-native speech. Most widely-adopted datasets, such as the Airline Travel Information Systems (ATIS) and the SNIPS Natural Language Understanding corpus, primarily consist of native speech samples, limiting the ability to accurately evaluate and compare the robustness of different approaches to non-native speech.

This work aims to investigate the difficulties posed by non-native speech with grammatical errors in SLU systems and shed light on the factors contributing to their failure in such instances. It conducts empirical experiments to comprehensively assess the performance of state-of-the-art NLU models when applied to Automatic Speech Recognition (ASR) transcripts of both native and non-native speech. The study focuses on evaluating the degradation of performance on critical tasks such as Intent Detection and Slot Filling, utilizing widely-adopted benchmark datasets like the Airline Travel Information Systems (ATIS) and the SNIPS Natural Language Understanding corpus. Furthermore, the research delves into exploring the significance of different Parts of Speech (POS) and conducting an in-depth analysis of the errors involving them. By examining the error rates and patterns associated with various POS categories, the study aims to identify the linguistic components that are particularly susceptible to errors when processing non-native speech, thereby providing valuable insights for future model improvements.

Based on the findings of this study, an augmentation strategy is proposed to enhance the performance of SLU systems in the context of non-native speech. This strategy involves strategically substituting words in the training corpus with alternatives possessing the same POS attributes, effectively simulating the grammatical errors and patterns commonly observed in non-native speech. By exposing the models to this augmented data during the training phase, the proposed approach aims to equip SLU systems with the necessary robustness and adaptability to effectively handle non-native speech inputs, ultimately improving their accuracy and reliability in real-world applications. By incorporating this augmentation strategy, SLU systems are expected to develop increased robustness and improved performance when processing non-native speech inputs, mitigating the impact of grammatical errors and accent-induced variations. Consequently, the augmented models should exhibit enhanced accuracy in tasks such as Intent Detection and Slot Filling, ultimately leading to more reliable and effective human-machine communication in real-world scenarios involving non-native speakers.

This thesis is organized into six chapters. Chapter 1 provides an overview of the problem, along with a brief summary of the current state of research in the domain. Chapter 2 presents a detailed review of the prevailing trends and methodologies in designing and building SLU systems, discussing the various approaches adopted by researchers and practitioners in the field.

Chapters 3 and 4 focus on the analysis of non-native speech, specifically examining the impact of accents and grammatical errors on SLU system performance. These chapters provide a comprehensive exploration of how these factors affect the robustness and accuracy of SLU models. Building on these insights, Chapter 5 proposes an augmentation strategy aimed at enhancing the resilience of SLU systems to non-native speech influences. Finally, Chapter 6 reflects on the overall findings of the thesis and suggests potential directions for future research in this area.

Chapter 2

Related Work

Intent detection and slot filling are two fundamental tasks in the field of spoken language understanding (SLU) that have garnered significant research attention in recent years. Accurate intent detection and slot filling are crucial for building robust conversational AI systems that can effectively interpret user utterances and extract relevant information to complete the intended tasks.

The task of intent detection, also referred to as intent classification or intent recognition, aims to determine the underlying intent or goal behind a user's utterance. Early work in this area leveraged traditional machine learning techniques, such as Support Vector Machines (SVMs) [6], Naive Bayes classifiers [12], and Decision Trees [36]. With the advent of deep learning, researchers have explored the use of Recurrent Neural Networks (RNNs) [26], Long Short-Term Memory (LSTM) networks [14], and Convolutional Neural Networks (CNNs) [17] for intent detection, leveraging their ability to capture sequential patterns and contextual information in utterances. More recently, transformer-based models like BERT [8] have achieved state-of-the-art performance on this task by effectively encoding utterances and learning rich contextualized representations.

The task of slot filling, also known as entity extraction or slot tagging, involves identifying and extracting specific pieces of information (slots or entities) from a user's utterance that are required to complete the intended task. Conditional Random Fields (CRFs) [16] have been a popular approach for slot filling, effectively modeling the sequential nature of the task. However, the rise of deep learning has led to the exploration of RNNs, LSTMs, and attention-based models for this task, demonstrating improved performance over traditional methods. Like intent detection, transformer-based models such as BERT have also been applied to slot filling, leveraging their ability to capture long-range dependencies and context.

Recognizing the interdependence between intent detection and slot filling, researchers have explored joint modeling approaches that leverage shared representations and jointly optimize both tasks [19]. This joint approach has been shown to improve overall performance by capturing the inherent relationships between intents and slots, enabling more accurate and efficient understanding of user utterances.

These tasks are central to the development of SLU systems, which are essential components of virtual assistants (e.g., Siri, Alexa, Google Assistant), chatbots, and conversational AI systems across

various domains, such as customer service, e-commerce, and task automation. As such, continuous research efforts are being made to improve the accuracy and robustness of intent detection and slot filling techniques, enabling more natural and effective human-computer interactions.

The transformer model has gained a lot of traction recently across a lot of different domains. It has proven to be effective across multiple domains on a large number of tasks, including but not limited to image classification, machine translation [3], [20], summarization and speech recognition [5]. It has also shown improved the performance of SLU and NLU systems.

The attention mechanism is a key component of this effectiveness. It enables the model to focus on the most relevant parts of the input data when producing an output, mimicking the human ability to selectively concentrate on certain aspects while disregarding others.

The attention mechanism was first introduced in the context of neural machine translation by [3]. In their encoder-decoder architecture for translation, the encoder maps the input sequence into a sequence of vectors, while the decoder generates the translated output sequence one word at a time. At each step of the decoder, an attention distribution over the encoded input sequence is computed to determine which portions of the input are most relevant for generating the next output word.

Formally, consider an encoder-decoder model where the encoder has transformed the input sequence $(x_1, x_2, ...x_n)$ into a sequence of vectors $(h_1, h_2, ...h_n)$. At each time step t in the decoder, the goal is to compute an attention distribution $\alpha_t = (\alpha_t^1, \alpha_t^2, ..., \alpha_t^n)$ over the encoded input vectors. This distribution indicates how much each input vector should be weighted when computing the context vector ct, which summarizes the relevant portions of the input sequence for the current decoding step:

$$c_t = \sum_i \alpha_t^i h_i$$

The context vector c_t is then used along with the previous decoder hidden state s_{t-1} to compute the current hidden state s_t and generate the output y_t. The attention distribution α_t is computed by the attention mechanism, which can take various forms. In [3], it uses a feedforward neural network that scores each (h_i, s_{t-1}) pair to indicate their relevance:

$$e_t^i = a(s_{t-1}, h_i)$$

$$\alpha_t^i = \frac{exp(e_t^i)}{\sum_j exp(e_t^j)}$$

where a is a feedforward neural network and t is normalized to sum to 1 using a softmax.

The attention mechanism has proved to be a powerful and flexible tool. It allows neural networks to selectively focus on and combine information from different locations, alleviating the need to encode everything into a fixed-length vector as was common in earlier architectures [4] [31]. This enables handling variable-length inputs more naturally and capturing long-range dependencies more effectively.

Different variations on the basic idea have been explored, such as multi-head attention used in the Transformer architecture [32]. Here, instead of computing a single attention distribution, the attention is

computed in multiple "heads" in parallel to allow the model to attend to different aspects of the input. The outputs of the heads are then concatenated:

$$\text{MultiHead(Q, K, V)} = \text{Concat}(head_1, ..., head_h)$$

where

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

This multi-head variant helps the attention learn complementary representations and improves performance. Self-attention is another important variant where the attention mechanism relates different positions of the same input sequence to compute representations of that sequence.

Many other extensions on the standard formulations have been explored, such as incorporating convolutional attentions [30], restricting the attention to particular regions [22], combining with convolutions [11], and more. Overall, the attention mechanism has emerged as a powerful concept that equips neural networks with dynamic, selective focusing capabilities leading to substantial accuracy gains across a wide spectrum of tasks.

In our methodology, we leverage the insights gained from attention weights to discern the specific components of the model inputs that play a significant role in the prediction process. Attention mechanisms enable us to identify which parts of the input data are accorded greater importance during the model's decision-making

In conclusion, the integration of transformer-based models and attention mechanisms has significantly advanced the field of SLU particularly in the tasks of intent detection and slot filling. The ability of transformers to capture long-range dependencies and context, coupled with the selective focusing capability of attention mechanisms, has enabled substantial improvements in the accuracy and robustness of conversational AI systems. These advancements have facilitated more natural and effective human-computer interactions,

Chapter 3

Effect of non-native accents on the performance of SLU Systems

3.1 Introduction

Spoken Language Understanding (SLU) has become increasingly popular in recent times due to its ability to enable natural human-machine communication. SLU systems play a critical role in accurately interpreting and processing language. However, a challenge arises when these systems encounter non-native accent. This research work seeks to explore the efficacy of SLU systems in handling speech containing non-native accent and shed light on the factors contributing to the failure of such systems in such instances.

SLU systems are predominantly designed to process and comprehend native speech, which is generated by individuals who are native speakers of a specific language. The success of SLU systems in understanding native speech can be attributed to several factors. One such factor is the utilization of extensive large-scale datasets during the training process [34]. These datasets often encompass a wide range of native speech examples, allowing the SLU systems to learn the patterns and structures inherent in native language usage. Additionally, the constituent modules within SLU systems, such as the Natural Language Understanding (NLU) module, are often equipped with abundant instances of native speech [2]. These modules serve as crucial components in the overall system, contributing to its effectiveness in recognizing and interpreting native speech. As a result, the statistical models employed by SLU systems are well-tailored to the characteristics and complexities of native speech, enabling them to perform optimally in those scenarios. However, these models may encounter challenges when confronted with non-native speech due to the accents that lie outside the training distributions of these models.

When non-native accents are introduced into the interaction, the performance of SLU systems tends to suffer notably. One of the main reasons for this is an increase in transcription errors, which can have a cascading effect on downstream tasks such as intent detection [33]. Non-native accents can introduce variations in pronunciation and speech patterns that deviate from what the SLU systems have been primarily trained on. This coupled with grammatical errors that are often made by non-native speakers, it becomes challenging for the models to accurately transcribe and interpret the provided utterance.

Previous research efforts in the domain have primarily concentrated on enhancing model performance by proposing novel architectures or addressing specific issues related to individual modules. Many studies have focused on developing new models or refining existing ones to improve overall accuracy and robustness. In their study [25], researchers propose an end-to-end SLU pipeline that aims to address and mitigate errors originating from the Automatic Speech Recognition (ASR) component. Their pipeline focuses on specifically targeting ASR errors to improve overall system performance. [28] introduces an augmentation strategy for training SLU models. This strategy involves simulating errors during the training process to enhance the model's resilience against variations and imperfections encountered in real-world scenarios. By exposing the model to synthetic errors, the authors aim to create a more robust system capable of effectively handling diverse speech patterns, accent variations, and common sources of noise in spoken language. Additionally, [29] proposes integrating a word confusion network into the SLU framework, which aids in mitigating errors by incorporating information about potential word confusions and their probabilities. These works primarily focus on refining architectures, proposing novel training strategies, and incorporating error mitigation techniques to improve the performance and robustness of SLU systems. However, they do not extensively address the intricacies and unique challenges of non-native language inputs. While these contributions advance SLU technology, they don't further our understanding of the intricacies of the failure of SLU systems with non-native speech.

This work aims to analyze the challenges presented by non-native speech in SLU systems. It conducts empirical experiments to evaluate the performance of NLU models when applied to ASR transcripts of non-native speech. Furthermore, it investigates the importance of various Parts of Speech (POS). The study primarily focuses on assessing the decline in performance of state-of-the-art models for Intent Detection and Slot Filling tasks, utilizing the ATIS [13] and SNIPS [7] datasets as benchmarks.

3.2 Methods

The construction of the Spoken Language Understanding (SLU) system in this study followed a cascade approach, which involved the integration of separate ASR and NLU modules. Fig. 3.1 provides an illustration for this approach. The ASR module converted the spoken input into textual transcriptions. To achieve accurate and robust speech recognition, we leveraged Whisper [24], a state-of-the-art ASR system known for its exceptional performance. Whisper has been widely recognized for its robustness in handling various acoustic conditions and challenging speech inputs.

3.2.1 NLU models

We consider three different state-of-the-art approaches for building the NLU models.

The first model (Bi-Model) used for the NLU module is a Bi-model based RNN semantic frame parsing network structure [35]. This model was specifically designed to perform joint intent detection and slot filling tasks, taking into account the cross-impact between these tasks. The model utilizes two

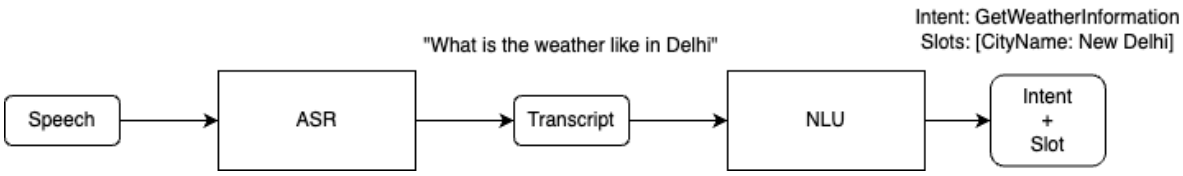

Figure 3.1 Cascaded SLU Design

correlated bidirectional LSTMs (BLSTM) to capture the inter-dependencies between intent detection and slot filling. The second model (Stack) is a novel framework for spoken language understanding (SLU) proposed by [23]. This framework aims to better incorporate intent information and guide the slot filling task, acknowledging the close relationship between intent and slots. In this framework, a joint model with Stack-Propagation is adopted, allowing the intent information to directly influence the slot filling process. By capturing the intent semantic knowledge, the model gains a deeper understanding of the underlying intent and leverages this information to enhance slot prediction. Furthermore, token-level intent detection is performed within the Stack-Propagation framework to alleviate error propagation and improve overall performance. The third model (SF-ID), proposed by [9], introduces a bi-directional interrelated architecture for joint intent detection and slot filling in SLU systems. It establishes direct connections between the tasks using a Slot Filling-Intent Detection network, enhancing their mutual influence. The model captures both local features and contextual dependencies.

3.2.2 Accented Speech data

Owing to the lack of standardized SLU datasets in different accents, we had to synthesize speech using Google's Text-to-Speech (TTS) service[1]. To evaluate the performance of the SLU systems with accented speech, we synthesized speech samples with American and Indian accents. This approach allowed us to simulate both native and non-native speech and assess how the SLU system handled variations in pronunciation and speech patterns associated with different accents. By introducing these accent variations, we aimed to gain valuable insights into the system's ability to effectively process and understand speech from diverse language backgrounds.

The use of synthesized speech samples provided us with control over the accent variations, ensuring consistency and reproducibility in our experiments. Google's TTS service, a widely-used and reliable tool allowed us to generate speech samples with American and Indian accents. This made it possible to simulate realistic scenarios where users with different accents interacted with the SLU system.

3.2.3 POS Analysis

To gain deeper insights into how different types of transcription errors influence the performance of NLU models, we categorized the utterances based on the part-of-speech (POS) category that was

[1] `https://cloud.google.com/text-to-speech`

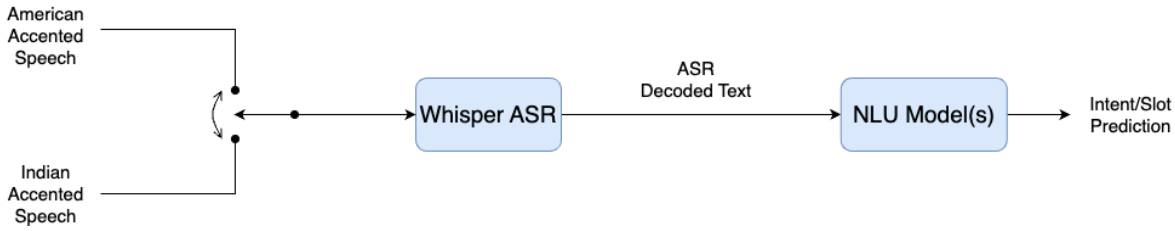

Figure 3.2 Accent Experiments Setup

corrupted by the error. This categorization enabled us to examine the impact of distinct transcription error types on the models' abilities to understand and correctly classify intents. By grouping the utterances according to the POS category that was modified, we could analyze how each specific error type affected the models' capabilities in understanding and accurately determining the intended meaning.

We compared the models' performance on the original, error-free versions of the utterances with their performance on the modified versions containing introduced errors within each POS group. This comparative analysis allowed us to identify patterns and trends, assessing which specific types of errors had the most significant impact on the models' accuracy. Additionally, we could evaluate whether certain POS categories were more challenging for the models to handle when errors were present. Through this analysis, we aimed to determine if the models exhibited heightened vulnerability to particular errors or if their performance degradation was consistent across various error categories.

3.3 Results

To assess the performance of the system in handling non-native speech with the proposed augmentation, we conducted evaluations focused on the NLU models. Firstly, we examined the performance of the NLU models on text with synthetically introduced grammatical errors. By evaluating the models' performance on such data, we aimed to understand their ability to handle and recover from grammatical inconsistencies. Furthermore, we sought to explore the impact of non-native accents on the system's performance. To achieve this, we utilized the transcripts of accented speech utterances and assessed the models' performance on these transcripts. By using speech data with non-native accents, such as Indian accent, we aimed to simulate scenarios involving speakers with different linguistic backgrounds. Through these evaluations, we aimed to gain insights into how the NLU models performed in the presence of non-native speech characteristics, including both grammatical errors and accents.

3.3.1 Model Performance Evaluation

We observed a degradation in model performance when testing with the ASR version compared to using gold transcripts, with a slightly worse performance for Indian accents in most cases compared to American accents. The results can be seen in Tables 3.1 and 3.2.

The drop in performance can be attributed to the inherent errors introduced by ASR systems during the transcription process. Non-native accents, such as Indian accents, pose additional challenges for ASR systems due to their distinct pronunciation patterns, intonation, and rhythm which are not part of the training data of the ASR models. These factors contribute to inaccuracies in transcriptions, leading to a slightly lower performance for Indian accents compared to American accents.

Table 3.1 NLU model performance on ATIS

	Intent Accuracy			**Slot F1Score**		
	Gold	American	Indian	Gold	American	Indian
SF-ID	95.6	94.83	93.94	97.89	97.43	97.19
Bi-Model	95.01	94.22	94.13	98.11	97.82	97.49
Stack	96.76	96.03	95.87	96.89	95.78	95.41

Table 3.2 NLU model performance on SNIPS

	Intent Accuracy			**Slot F1Score**		
	Gold	American	Indian	Gold	American	Indian
SF-ID	96.78	96.71	96.13	93.9	93.81	92.6
Bi-Model	97.55	97.25	97.06	95.47	95.21	94.74
Stack	97.09	96.46	95.26	95.37	94.88	93.82

3.3.2 Effect of Transcription Errors on NLU Performance

Figures 3.3 and 3.4 show the performance of the NLU model with increase in transcription errors. We observe a consistent decline in model performance as the number of transcription errors increased in the ASR output. Both Intent Detection and Slot Filling tasks exhibited a steady decrease in performance as the number of mistranscribed tokens grew in the sequence passed to the NLU model.

These transcription errors had a significant impact, introducing incorrect information into the input sequence and leading to confusion and misinterpretation by the NLU model. Consequently, the model struggled to accurately detect user intents and populate the appropriate slots with relevant information.

3.3.3 POS Analysis

Figure 3.5 displays the increase in error rates pertaining to the task of intent detection across both the ATIS and SNIPS datasets. Each data point represents the percentage of instances where a transcription error associated with a specific POS led to a misclassification of intent by the model(s). Certain POS, particularly Nouns and Auxiliary Verbs, consistently exhibit elevated error rates across all three models and datasets. This consistent pattern underscores the significance of these POS in determining the intent accurately.

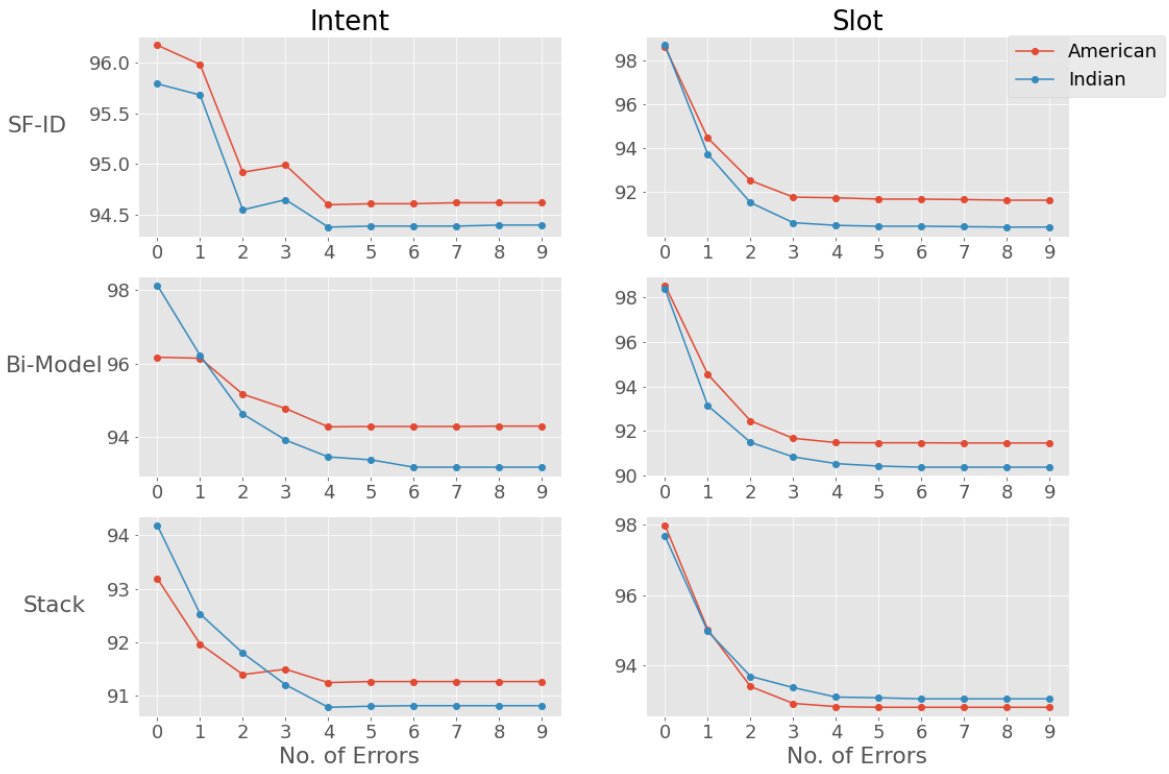

Figure 3.3 Performance on ATIS with Cumulative Transcription Errors

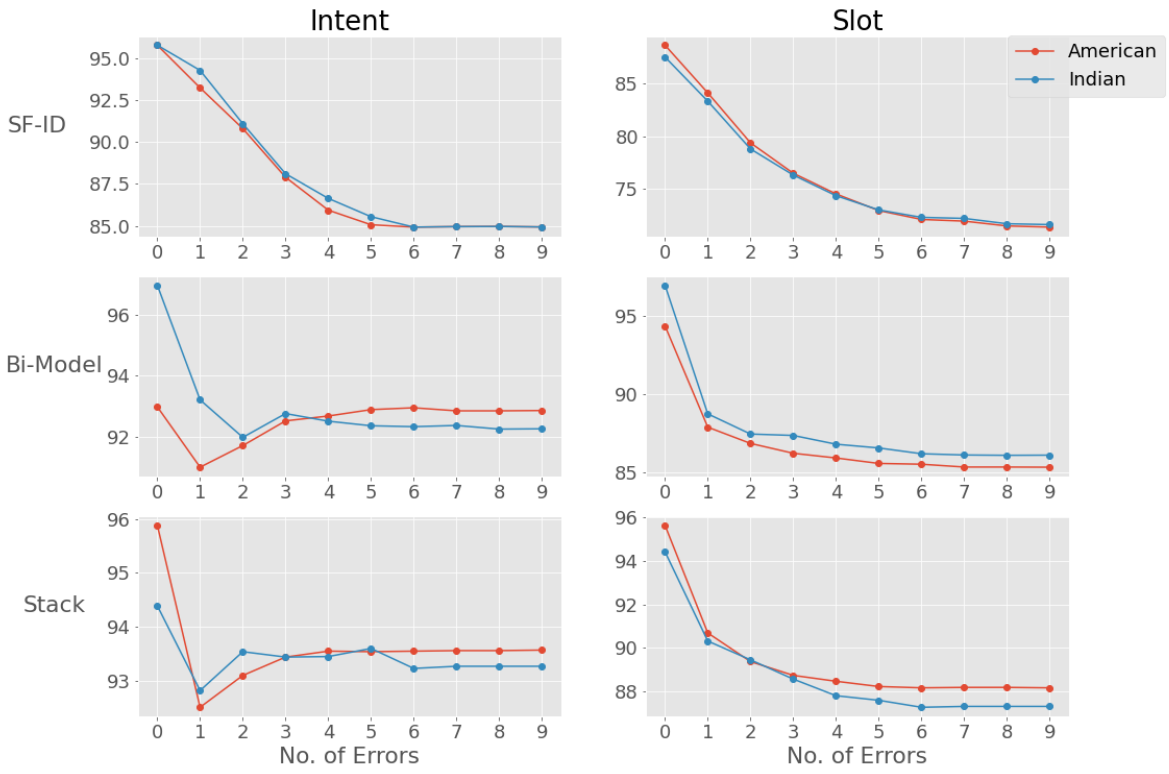

Figure 3.4 Performance on SNIPS with Cumulative Transcription Errors

Figure 3.5 Per POS increase in Intent Error Rates

3.4 Conclusion

This work underscores the necessity of addressing the unique challenges posed by non-native speech in SLU systems. While SLU systems perform well with native speech due to extensive training on large-scale datasets, they struggle with non-native accents, leading to increased transcription errors and a decline in intent detection and slot filling accuracy. Our empirical analysis reveals a performance drop when models encounter non-native speech, particularly highlighting the compounded difficulties introduced by transcription errors and variations in pronunciation. By leveraging state-of-the-art ASR and NLU models and synthesizing speech data with controlled accent variations, we provide a comprehensive evaluation of current SLU systems' robustness. These findings highlight the critical need for more inclusive and diverse training data, as well as advanced models capable of adapting to the linguistic diversity inherent in real-world applications.

Chapter 4

Effect of Grammatical Errors on the performance of SLU Systems

4.1 Introduction

Spoken Language Understanding (SLU) have gained significant popularity in recent years, as it provides a way for humans to interact with machines in a natural mode of communication. SLU systems are crucial as they are responsible for accurately interpreting and processing language. SLU systems play a pivotal role in various tools, such as digital assistants, and are extensively utilized by individuals not only in their native languages but also in non-native languages [21]. Consequently, these SLU systems must be capable of effectively process and comprehend non-native characteristics of speech. However, SLU systems have a limitation in working with speech that contains non-native grammatical errors [37]. This research paper aims to investigate the effectiveness of SLU systems in processing speech that contains non-native grammatical errors and highlight the reasons why SLU systems fail in such cases.

SLU systems are primarily designed to recognize and interpret native speech, which is produced by individuals who are native speakers of a particular language. The effectiveness of SLU systems in recognizing native speech can be attributed to the large datasets used to train them [34], or the constituent modules like [2] which contain ample examples of native speech. Consequently, the statistical models used by these systems are better suited to the patterns and structures of native speech.

Non-native speakers of a language often make errors, which can include pronunciation, grammar, or syntax errors. These errors are often caused by the influence of the speaker's native language, which can have different patterns and structures [18]. Non-native grammatical errors pose a significant challenge for SLU systems because the statistical models used by these systems are not equipped to handle the varied patterns and structures of non-native speech. When non-native grammatical errors are introduced into speech, the effectiveness of SLU systems is significantly reduced. The errors can lead to incorrect interpretations of the spoken language, resulting in inaccurate transcriptions or misinterpretations of the speaker's intent. Additionally, non-native errors can cause ASR systems to misinterpret common words, resulting in a failure to recognize the intended meaning of an utterance [33].

SLU systems, which are the core component of interfaces like Dialog Systems, have a limitation in processing non-native speech. The challenges posed by non-native grammatical errors can lead to

incorrect interpretations, resulting in inaccurate transcriptions and by extension misinterpretations of the speaker's intent. The investigation of the effectiveness of SLU systems in processing speech that contains non-native grammatical errors is crucial to improving speech technology's accuracy.

Previous research efforts in the domain have primarily concentrated on enhancing model performance by proposing novel architectures or addressing specific issues related to individual modules like noisy environments [1] or large scale NLU. Many studies have focused on developing new models or refining existing ones to improve overall accuracy and robustness. In [25], the researchers propose an end-to-end SLU pipeline. This pipeline is designed to specifically target and mitigate errors stemming from the ASR component. The authors of [28] introduce an augmentation strategy for training SLU models. This strategy involves simulating errors during the training process to enhance the model's robustness against variations and imperfections commonly encountered in real-world scenarios. By exposing the model to synthetic errors, the authors aim to create a more resilient system that can effectively handle diverse speech patterns, accent variations, and other sources of noise commonly encountered in spoken language. In [29], integration of a word confusion network is proposed into the SLU framework. This integration allows for effective error mitigation by incorporating information about potential word confusions and their likelihoods. The focus of these works lies predominantly on refining architectures, proposing novel training strategies, or incorporating error mitigation techniques, rather than delving into the intricacies and unique challenges posed by non-native language inputs.

This work aims to explore the effect of non-native speech on SLU systems. We evaluate the performance of NLU models on input with grammatical errors, to explore the degradation of performance of State of the Art models on the tasks of Intent Detection and Slot Filling on the Airline Travel Information Systems (ATIS) and SNIPS Natural Language Understanding benchmark datasets. We also conduct an analysis of the errors made by SLU systems when presented with grammatically erroneous input. Specifically, we examine the error rates with respect to the different Parts of Speech (POS) for the task of Intent Detection. We found that the performance of State of the Art NLU models drop considerably when presented with inputs that had L2 learner inspired grammatical errors. We also found certain POS to be more important, corruption of which had a higher chance of resulting in an error.

4.2 Methods

Table 4.1 Dataset Statistics

Dataset	Train Size	Test Size	# Intent Classes	# Slot Classes	Mean # Tokens per Utterance
ATIS	4478	893	21	120	12.28
SNIPS	13084	700	7	72	10.0

To assess the impact of grammatical errors on the performance of Natural Language Understanding (NLU) models, we conducted experiments using three models, proposed in [35], [23] and [9]. These

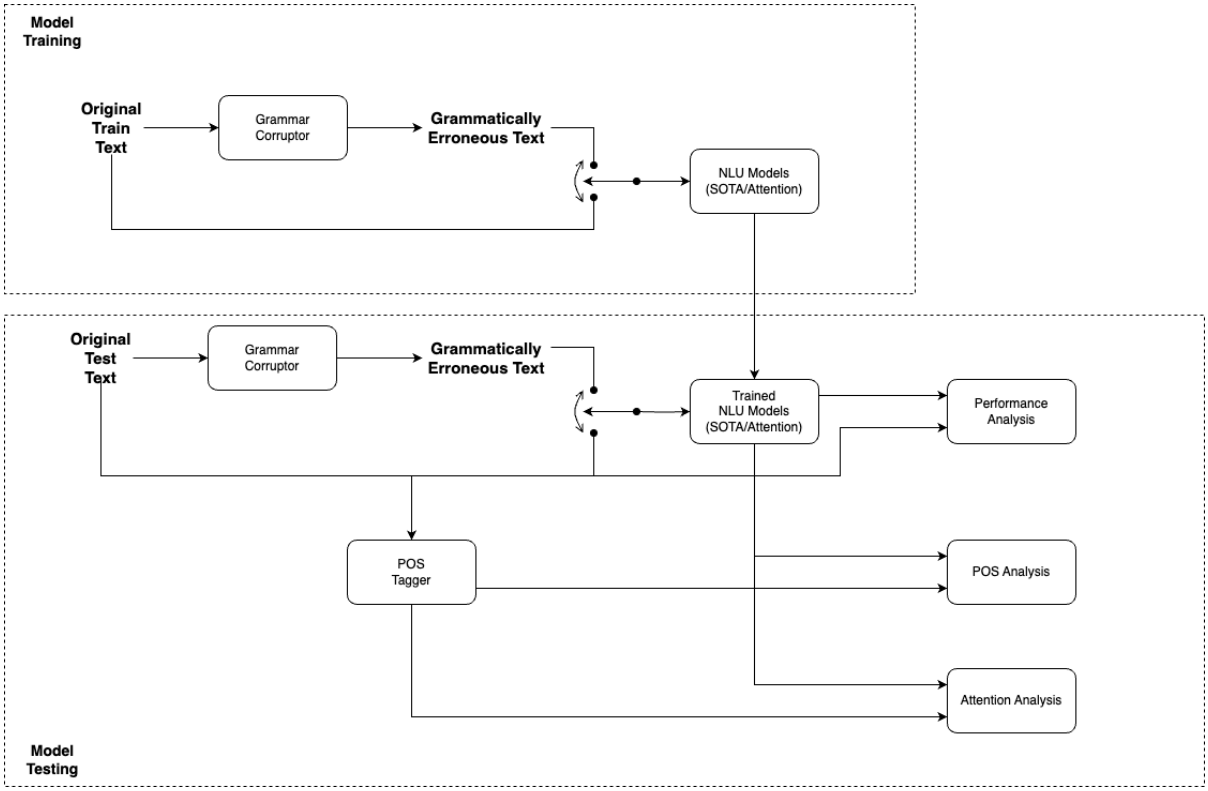

Figure 4.1 Workflow for the conducted analysis

models were chosen based on their State of the Art performance on the ATIS and SNIPS benchmarks. These are referred to as SF-ID-Network-For-NLU (SF-ID), Bi-Model-Intent-And-Slot (Bi-Model) and StackPropagation-SLU (Stack) respectively. By evaluating these models in the presence of grammatical errors, we aimed to understand how robust they are in handling linguistic variations and to gauge their overall performance in real-world scenarios.

The ATIS [13] and SNIPS [7] benchmarks provide standardized datasets that are widely used for evaluating the performance of NLU models. These datasets include utterances that cover various intents and are representative of the tasks performed in the respective domains. Statistics for the datasets can be found in Table 4.1.

Figure 4.1 presents a comprehensive overview of the analysis workflow, which was undertaken to evaluate the impact of grammatical errors on the performance of the models. The initial step of this study involved training two distinct versions of each model, with one version trained on the original, error-free text from the dataset(s) and the other on the grammatically erroneous version of the same text.

Following the model training, the test set was utilized to assess the performance of each model variant. To gain further insights into the impact of errors on the model's performance, an additional analysis was conducted by passing the data through a Part of Speech (POS) tagger in the spaCy toolkit [15]. This POS information allowed us to categorize the errors based on their corresponding parts of speech, such as nouns, verbs, adjectives, and so forth. Evaluating the model's performance along different POS provided valuable information about which linguistic elements were more susceptible to errors and which parts of speech might have a greater influence on the overall performance of the models.

4.2.1 Synthesizing Grammatical Errors

We utilized the tool introduced in [10] to synthesize grammatically erroneous sentences by introducing grammatical errors. The tool offered four operations: insert, delete, substitute, and move, enabling us to create random syntactic noise and change word forms. We also removed quotes from the corpus to generate additional variations. For a given input sentence, the tool applies the above stated operations and introduces noise. By utilizing POS information, the tool has the capability to replace words with incongruent versions. This means that words can be replaced with others that might not belong to the same POS category, allowing us to study the impact of a wide variety of grammatical errors on the performance of NLU tasks. This is particularly crucial as real-world non-native language usage often involves instances where words may be used in unconventional ways.

4.2.2 Intent Detection and Slot Filling Performance

In our experiments, we trained the three models using the regular train splits available in the dataset. The models were trained to learn the relationships between input utterances and their corresponding intents and slots based on this clean and error-free training data.

To evaluate the impact of grammatical errors on the performance of the models, we conducted testing on two versions of the test splits. The first version is the original test split, which contains naturally occurring utterances without any introduced errors. This allows us to assess the models' performance under normal conditions and compare it to their performance on the clean training data. The second version of the test split was included modifications for synthetic grammatical errors. These errors were introduced by corrupting specific parts-of-speech (POS) in the utterances, simulating common grammatical mistakes that users might make in real-world interactions. By testing the models on this modified split, we aimed to evaluate their ability to handle and understand utterances with grammatical errors, which are more representative of the variability and challenges encountered in real-world language understanding scenarios.

By comparing the models' performance on the original test split versus the modified split with introduced grammatical errors, we can gain insights into their robustness and adaptability to linguistic variations. To explore strategies for mitigating the impact of grammatical errors on NLU model performance, we trained the models on a modified version of the training data. By introducing synthetic grammatical errors into the original training data, we aimed to expose the models to linguistic variations and errors commonly found in real-world user inputs. This training approach sought to enhance the models' ability to handle and adapt to grammatical errors.

4.2.3 POS Analysis

To gain a deeper understanding of the effects of specific kinds of grammatical errors on the performance of the NLU models, we categorized the utterances based on the type of part-of-speech (POS) that was intentionally corrupted to introduce the error. This categorization allowed us to examine the impact of different types of grammatical errors on the models' performance.

By grouping the utterances based on the POS that was modified, we were able to analyze how each specific type of error affected the models' ability to understand and classify the intent correctly. Comparing the performance of the models on the original, error-free versions of the utterances to their performance on the modified versions with introduced errors within each group of POS, allows us to identify patterns and trends. This analysis enabled us to assess which specific types of grammatical errors had the most significant impact on the models' accuracy and whether certain POS categories were more challenging for the models to handle.

4.2.4 Attention Analysis

As an alternative approach to analyzing the importance of specific POS for intent classification, we trained an attention-based NLU model on the task of intent classification. By leveraging this attention mechanism, we could explore the relative importance of different parts-of-speech for each intent class in a more nuanced and interpretable manner.

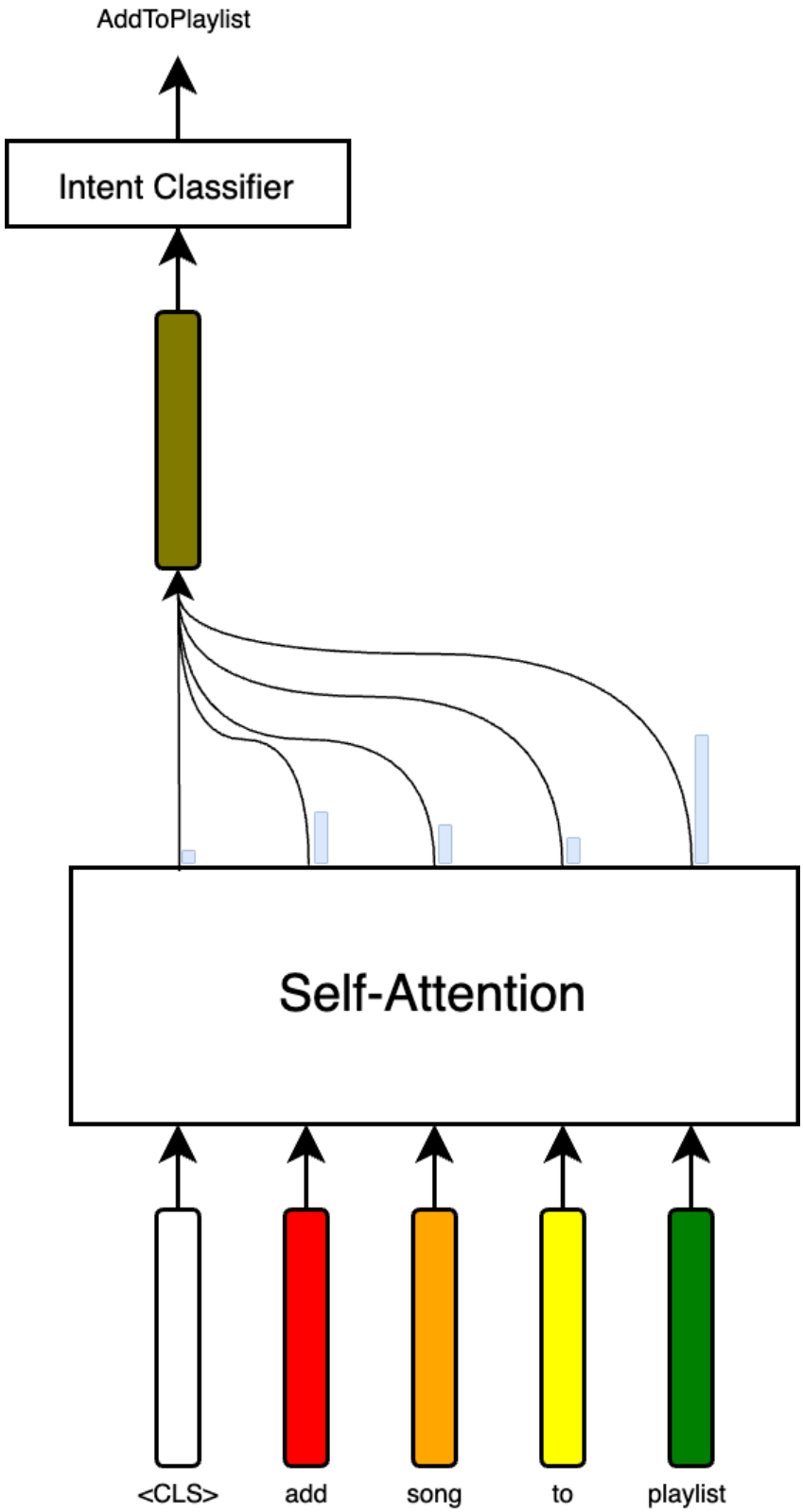

Figure 4.2 How attention mechanism weights information from different tokens

Table 4.2 NLU model performance on ATIS and SNIPS with and without grammatical errors

			ATIS			SNIPS		
			SF-ID	**Bi-Model**	**Stack**	**SF-ID**	**Bi-Model**	**Stack**
Intent Accuracy	O-Train	O-Test	95.63	95.18	96.19	97.00	97.42	97.71
		E-Test	95.18	94.40	93.50	95.85	92.71	96.14
	E-Train	O-Test	96.19	95.40	93.72	97.00	96.28	96.85
		E-Test	96.19	95.63	94.28	97.14	96.14	97.42
Slot F1 Score	O-Train	O-Test	97.94	97.81	97.45	94.00	95.32	95.40
		E-Test	97.15	97.09	96.66	90.87	90.21	92.24
	E-Train	O-Test	97.82	97.73	97.40	92.46	93.18	93.79
		E-Test	97.90	97.70	97.50	93.53	94.28	95.08

The attention-based NLU model was designed to take sequences of words as input and produce a probability distribution over the set of possible intent classes. During the training process, the model learned to assign higher attention weights to the parts of the input that were most relevant for determining the correct intent class. By analyzing these learned attention weight distributions, we could gain insights into which parts-of-speech carried the most discriminative information.

Concretely, after training the attention model, we could examine the attention weights assigned to words belonging to different parts-of-speech categories, such as nouns, verbs, adjectives, and so on. By aggregating these attention weights across examples, we could determine which POS received consistently higher weights by the model. This analysis allowed us to assess the relative importance and contribution of different parts-of-speech toward accurately recognizing intent classes, providing a more fine-grained understanding of the linguistic patterns and cues leveraged by the model.

4.3 Results

4.3.1 Intent Detection and Slot Filling

Table 4.2 provide insights about the performance of the NLU models on the ATIS and SNIPS datasets for the tasks of Intent Detection and Slot Filling. The table demonstrate the performance of the models across different variations of the train and test splits. Notably, there is a clear drop in performance for the models trained on the original train split (O_train) when transitioning from testing on the original test split (O_test) to the test split with introduced grammatical errors (E_test). This drop indicates the challenges posed by grammatical errors in real-world language understanding scenarios.

The results obtained from training the models on a modified version of the train split (E_train) indicate a potential mitigation strategy for addressing the drop in model performance caused by grammatical

Table 4.3 Cumulative Intent Accuracy and Slot F1-Score with Increased Errors on ATIS and SNIPS

	ATIS						SNIPS					
	SF-ID		Bi-Model		Stack		SF-ID		Bi-Model		Stack	
# Corrupted	Intent	Slot	Intent	Slot	Intent	Slot	Intent	Slot	Intent	Slot	Intent	Slot
0	95.28	98.47	94.88	98.05	94.86	97.41	97.64	96.10	94.44	95.75	96.32	94.58
1	96.11	97.83	95.26	97.57	94.53	96.85	96.26	93.68	93.42	93.81	96.25	93.93
2	95.60	97.47	94.95	97.29	94.13	96.92	96.15	92.25	93.63	91.70	96.44	92.80
3	95.44	97.42	94.74	97.29	93.72	96.77	96.06	91.14	92.57	90.73	96.08	92.32
4	95.35	97.26	94.66	97.19	93.47	96.73	96.02	90.89	92.76	90.39	96.13	92.28
5	95.17	97.18	94.38	97.12	93.49	96.70	95.82	90.92	92.64	90.22	96.14	92.24
6	95.17	97.17	94.39	97.10	93.50	96.66	95.85	90.91	92.70	90.25	96.14	92.24
7	95.17	97.17	94.39	97.10	93.50	96.66	95.85	90.91	92.70	90.25	96.14	92.24
8	95.17	97.17	94.39	97.10	93.50	96.66	95.86	90.87	92.71	90.21	96.14	92.24
9	95.18	97.15	94.40	97.09	93.50	96.66	95.86	90.87	92.71	90.21	96.14	92.24

errors. By training the models on a train split that includes synthetic grammatical errors, we observe a recovery of a significant portion of the performance loss when tested on the modified test split with introduced errors. For example, the Stack model lost 2.69% Intent Accuracy on ATIS when tested on grammatically erroneous version but training on synthetic dataset allowed us to recover 0.78%.

Table 4.3 provides a quantitative understanding of how the models' performance deteriorates as the number of corrupted tokens (# Corrupted) increases. This tells us that most models become progressively worse as the tokens are replaced and information available is diminished which makes the task more challenging. This information helps us gauge the sensitivity of the models to different levels of grammatical errors and provides insights into their robustness and generalization capabilities.

4.3.2 POS Analysis

Table 4.4 presents the results of POS analysis experiments. It shows the drop in performance with respect to POS that were corrupted in the input sentence. A larger value indicates a larger effect. This provides us with valuable insights into the impact of specific POS on the performance of the model for the task of Intent Detection.

We can see that certain POS like Auxiliary Verbs and Noun had a consistently large effect across the three models and both of the datasets.

4.3.3 Attention Analysis

The results presented in Table 4.5 provide insights into the attention analysis conducted in our study. It demonstrates the average attention weights assigned to different parts of speech (POS) during the intent classification task. A higher value represents indicates that more attention was paid to that particular

Table 4.4 Increase in Error Rates per POS when tested on Grammatically Erroneous data for ATIS and SNIPS

	ATIS			SNIPS		
POS	**SF-ID**	**Bi-Model**	**Stack**	**SF-ID**	**Bi-Model**	**Stack**
ADJ	0.00	0.00	0.00	0.02	0.06	0.00
AUX	0.02	0.04	0.04	0.01	0.04	0.04
DET	0.01	0.00	0.02	0.00	0.07	0.02
NOUN	0.03	0.11	0.03	0.03	0.05	0.03
NUM	-0.01	0.00	0.00	0.02	0.03	0.00
PRON	0.00	0.00	0.00	0.00	0.00	0.2
PROPN	0.01	-0.01	0.01	-0.01	0.05	0.00
VERB	0.00	0.00	0.00	0.03	0.09	0.04

POS which indicates that it played an important role in the classification decision of the model. We can observe that certain POS like Proper Noun and Adposition (Prepositions + Postpositions) receive higher attention weights, for both datasets, indicating their perceived importance in determining the intent of a given utterance. Please Note that X refers to 'other' which are tokens that couldn't be tagged into any of the presented POS.

4.4 Conclusion

In conclusion, this chapter thoroughly examined the challenges posed by non-native grammatical errors to SLU systems. Through evaluation on benchmark datasets like ATIS and SNIPS, it has demonstrated significant performance degradation of state-of-the-art NLU models when confronted with such errors. Analysis of error patterns across different Parts of Speech (POS) has underscored the importance of linguistic components in SLU accuracy, particularly in non-native contexts. These findings emphasize the critical need for advancing SLU technologies to effectively handle diverse linguistic variations, thereby enhancing system robustness and reliability in practical applications.

Table 4.5 Mean Attention Weights per POS on ATIS and SNIPS

POS	**ATIS**	**ATIS_freq**	**SNIPS**	**SNIPS_freq**
ADJ	0.06	288	0.06	349
ADP	0.12	1773	0.04	891
ADV	0.04	76	0.10	92
AUX	0.05	328	0.05	280
CCONJ	0.01	135	0.04	49
DET	0.04	583	0.04	683
INTJ	0.01	51	0.07	35
NOUN	0.04	1878	0.09	1770
NUM	0.05	211	0.05	285
PART	0.19	191	0.04	95
PRON	0.10	586	0.05	475
PROPN	0.10	2198	0.19	493
SCONJ	0.07	34	0.09	37
VERB	0.05	932	0.13	832
X	0.33	7	0.33	6
SPACE	1	0.13	0.00	1

Chapter 5

POS based augmentation for L2 learner based Errors

5.1 Introduction

Natural Language Understanding (NLU) is a critical component of intelligent systems, enabling them to comprehend and interpret human language. However, the performance of NLU models is often hindered by the limited diversity and size of training data. To address this challenge, researchers have explored various data augmentation techniques, aiming to expand the corpus and enhance the robustness of these models. In this study, we propose a novel approach to text augmentation, leveraging the systematic substitution of words based on their parts-of-speech (POS) attributes. We focus on Proper Nouns, leveraging SpaCy for named entity recognition, to categorize and replace tokens.

The proposed methodology revolves around the concept of substituting words in the corpus with alternative lexicons that share the same POS properties. By systematically replacing tokens with their POS-based counterparts, we can significantly expand the corpus while preserving the linguistic structure and semantic integrity of the original text. This approach not only increases the diversity of the training data but also exposes the NLU models to a broader range of linguistic variations, potentially enhancing their generalization capabilities and overall performance.

To validate the effectiveness of our approach, we conducted experiments on two widely-used NLU tasks: intent classification and slot detection. By employing POS-based text augmentation during model training, we observed improvements in performance across various models and datasets. Notably,

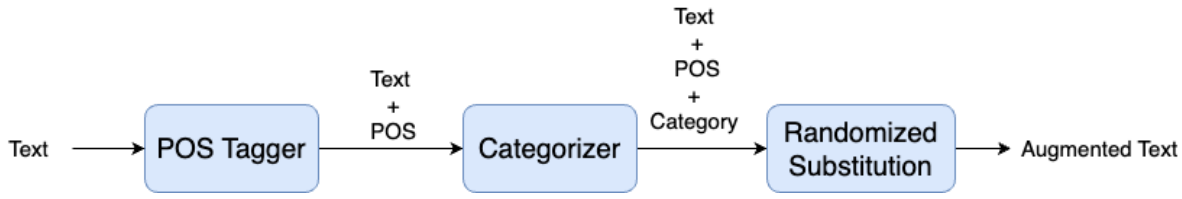

Figure 5.1 Augmentation Pipeline

the results demonstrated that certain models, such as the Bi-Model and SF-ID, exhibited substantial performance gains, particularly in the context of intent classification and slot detection tasks, respectively.

5.2 Methods

Based on the findings in the preceding analysis, we proceed to enhance the efficacy of natural language understanding (NLU) models through the employment of text augmentation techniques. We propose a methodology where words in the corpus are substituted with alternatives that have the same parts-of-speech (POS) attributes. By systematically substituting lexicons based on shared POS properties, we can significantly expand the corpus.

In order to generate a comprehensive pool of tokens for replacement and dataset augmentation, the initial step involves assigning a part-of-speech (POS) tag to each token within the corpus. Subsequently, tokens belonging to the specific POS category under consideration—in this instance, Proper Nouns—are selected. These tokens are then categorized based on various criteria. It's noteworthy that the methodology employed in this module can vary depending on the POS under examination. For instance, in the case of verbs, leveraging resources such as VerbNet [27] can be used. We experimented with Proper Nouns, hence used Named Entity Recognition provided by SpaCy[1]. Ultimately, to produce the augmented version of the dataset, each token (specifically entities in this context) is substituted with alternatives sourced from the corresponding category pool. This systematic approach enables the generation of an expanded dataset, facilitating more comprehensive model training.

5.3 Results

Tables 5.1 and 5.2 present the results of incorporating our proposed part-of-speech (POS) based augmentation technique during the training of various natural language understanding models. Across both the intent classification and slot detection tasks, on the ATIS and SNIPS datasets, we observe a noticeable improvement in performance for most of the evaluated models when compared to their performance without data augmentation.

Particularly noteworthy is the substantial performance gain achieved by the Bi-Model architecture on the intent classification task, exhibiting the highest jump in accuracy among all models tested on both the ATIS and SNIPS datasets. This significant improvement highlights the effectiveness of our POS-based augmentation approach in enhancing the Bi-Model's ability to comprehend and accurately classify user intents, even in the presence of linguistic variations and diverse phrasing.

For the slot detection task, the models that demonstrated the most substantial performance improvement were the SF-ID and Bi-Model architectures. Specifically, on the ATIS dataset, the SF-ID model exhibited the largest accuracy gain, while the Bi-Model showed the most significant performance boost

[1] `https://spacy.io/api/entityrecognizer`

on the SNIPS dataset. These findings suggest that our augmentation technique effectively expanded the training data with diverse linguistic patterns, enabling these models to better capture and generalize the underlying relationships between utterances and their corresponding slot labels.

Table 5.1 Model performance on ATIS on erroneous version of test split

	Intent Accuracy		**Slot F1Score**	
	w/o augmentation	with augmentation	w/o augmentation	with augmentation
SF-ID	94.78	95.31	96.99	97.36
Bi-Model	94.57	95.55	97.29	97.42
Stack	93.39	92.71	96.23	96.2

Table 5.2 Model performance on SNIPS on erroneous version of test split

	Intent Accuracy		**Slot F1Score**	
	w/o augmentation	with augmentation	w/o augmentation	with augmentation
SF-ID	95.83	95.35	91.19	91.56
Bi-Model	92.52	92.83	90.79	91.29
Stack	96.55	97.28	92.13	92.35

Table 5.3 Model performance with and without augmentation during training on ATIS

	Intent Accuracy						**Slot F1-Score**					
	Gold		American		Indian		Gold		American		Indian	
	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug
SF-ID	94.39	96.1	94.66	94.93	94.95	93.84	96.71	97.41	97.01	92.33	96.86	91.95
Bi-Model	93.87	95.67	93.93	93.39	94.18	93.81	96.89	97.89	96.96	91.36	96.98	90.58
Stack	94.59	92.91	94.98	90.96	94.78	91.05	95.92	97.24	96.13	92.83	95.94	93.75

Table 5.4 Model performance with and without augmentation during training on ATIS

	Intent Accuracy						**Slot F1-Score**					
	Gold		American		Indian		Gold		American		Indian	
	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug	no-aug	aug
SF-ID	95.52	95.61	95.78	84.94	96.1	84.99	93.32	92.16	93.32	71.36	93.05	71.03
Bi-Model	96.47	96.53	96.72	93.19	96.64	91.55	94.26	93.11	94.26	85.28	94.18	86.03
Stack	96.22	96.27	96.53	93.3	96.43	92.31	94.45	93.57	94.45	87.14	94.43	86.68

5.4 Conclusion

The chapter investigated the efficacy of a novel approach to text augmentation in improving NLU models' performance. By systematically substituting tokens based on their POS attributes, particularly focusing on Proper Nouns identified through SpaCy's named entity recognition, we aimed to enhance model robustness and generalization capabilities. Experimental results across intent classification and slot detection tasks on ATIS and SNIPS datasets demonstrated significant performance gains for various models, notably the Bi-Model architecture, showcasing its enhanced accuracy and slot F1-score compared to unaugmented counterparts. These findings underscore the potential of POS-based augmentation techniques to expand training data diversity effectively, thereby improving NLU model performance in handling linguistic variations and diverse phrasing in practical applications.

Chapter 6

Conclusion and Future Works

In this work, we undertook a comprehensive investigation into the performance limitations of spoken language understanding (SLU) systems when confronted with non-native accents and grammatical errors. By constructing SLU pipelines incorporating state-of-the-art automatic speech recognition and natural language understanding models, we were able to benchmark their performance on standard datasets and isolate the impact of accented speech and ungrammatical utterances through synthetic data generation.

Our empirical evaluation revealed significant performance degradation when SLU models encountered non-native accents and grammatical deviations, highlighting the need for targeted strategies to enhance their robustness and generalization capabilities. The attention-based architectures employed in our study further corroborated these findings, shedding light on the vulnerabilities of current models with regards to reliance on particular Parts of Speech (POS) like Proper Nouns and Auxiliary Verbs.

Leveraging these insights, we proposed a novel data augmentation strategy specifically designed to address the performance bottlenecks associated with non-native speech. By systematically introducing controlled variations in accent and grammatical errors during training, our approach aims to improve the models' ability to handle such challenges, ultimately paving the way for more inclusive and accessible SLU systems.

While our proposed augmentation provides some recovery of the lost performance, further research is needed. Exploring more advanced techniques for synthetically generating accented and ungrammatical speech data could lead to even more effective augmentation approaches. Additionally, investigating the impact of different types of accents and grammatical errors could provide valuable insights into the specific vulnerabilities of SLU models, informing targeted architectural enhancements or fine-tuning strategies.

An avenue that warrants further exploration is the impact of different automatic speech recognition (ASR) models on the overall SLU pipeline performance for non-native speech. In our current study, we utilized a single, fixed-size ASR model. However, evaluating a diverse range of ASR architectures, including larger models with increased capacity or specialized models tailored for accented speech recognition, could provide valuable insights. Such an analysis could help delineate the specific errors and

limitations introduced by the ASR component, and how those propagate and influence the downstream natural language understanding tasks.

Lastly, as SLU systems become increasingly deployed in real-world scenarios, it is crucial to study their performance and adaptability across diverse domains and applications. Developing domain-adaptation techniques and transfer learning approaches could further enhance the generalization capabilities of these models, ensuring their effectiveness in a wide range of practical settings, including those involving non-native speakers.

Related Publications

1. **Ranjan, S.**, Nanduri, S.K., Virdi, P., Yarra, C. (2023). Analysis of Natural Language Understanding Systems with L2 Learner Specific Synthetic Grammatical Errors Based on Parts-of-Speech. In SPECOM 2023. Lecture Notes in Computer Science(), vol 14338. Springer, Cham. https://doi.org/10.1007/978-3-031-48309-7_36

Bibliography

[1] M. N. Ali, V. J. Schmalz, A. Brutti, and D. Falavigna. A speech enhancement front-end for intent classification in noisy environments. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 471–475, 2021.

[2] D. Amodei et al. Deep speech 2 : End-to-end speech recognition in english and mandarin. volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[5] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 577–585, Cambridge, MA, USA, 2015. MIT Press.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190, 2018.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] H. E, P. Niu, Z. Chen, and M. Song. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] J. Foster and O. Andersen. GenERRate: Generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[11] J. Gehring, M. Auli, D. Grangier, and Y. Dauphin. A convolutional encoder model for neural machine translation. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[12] D. J. Hand and K. Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review / Revue Internationale de Statistique*, 69(3):385–398, 2001.

[13] C. T. Hemphill, J. J. Godfrey, and G. 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.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[15] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[17] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[18] J. Lee and S. Seneff. An analysis of grammatical errors in non-native speech in english. In *2008 IEEE Spoken Language Technology Workshop*, pages 89–92, 2008.

[19] B. Liu and I. Lane. Joint online spoken language understanding and language modeling with recurrent neural networks. In R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, and A. Gainer, editors, *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 22–30, Los Angeles, Sept. 2016. Association for Computational Linguistics.

[20] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[21] D. Pal, C. Arpnikanondt, S. Funilkul, and V. Varadarajan. User experience with smart voice assistants: The accent perspective. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2019.

[22] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[23] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2078–2087, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[25] M. Rao, A. Raju, P. Dheram, B. Bui, and A. Rastrow. Speech to semantics: Improve asr and nlu jointly via all-neural interfaces. *ArXiv*, abs/2008.06173, 2020.

[26] D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. 1987.

[27] K. K. Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006.

[28] E. Simonnet, S. Ghannay, N. Camelin, and Y. Estève. Simulating ASR errors for training SLU systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[29] E. Simonnet, S. Ghannay, N. Camelin, Y. Estève, and R. de Mori. ASR error management for improving spoken language understanding. In *Interspeech 2017*, Stockholm, Sweden, Aug. 2017.

[30] S. Sukhbaatar, E. Grave, P. Bojanowski, and A. Joulin. Adaptive attention span in transformers. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy, July 2019. Association for Computational Linguistics.

[31] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] N. T. Vu, Y. Wang, M. Klose, Z. Mihaylova, and T. Schultz. Improving asr performance on non-native speech using multilingual and crosslingual information. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[34] P. Wang, L. Wei, Y. Cao, J. Xie, and Z. Nie. Large-scale unsupervised pre-training for end-to-end spoken language understanding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7999–8003, 2020.

[35] Y. Wang, Y. Shen, and H. Jin. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 309–314, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[36] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[37] F. Yin, Q. Long, T. Meng, and K.-W. Chang. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online, July 2020. Association for Computational Linguistics.