International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 13 Issue 5, May-2024, Impact Factor: 8.375

Translation Management Model: An Integrated ASR and MT Model (Speech to Text Translation) using LLM

Aviral Mehlawat

Amity School of Engineering Technology, Amity University, Haryana, India

ABSTRACT

In terms of Development of Artificial Intelligence and Machine Learning, everyday development of new technologies enhances the speed of development. When talked of ASR and MT used together, the concept of integrating the ASR and the MT models were the revolutionary initiatives by the researchers at IBM (Brown et al., 1994) and the TransTalk project (Dymetman et al., 1994; Brousseau et al., 1995). Evolution with time and deployment of MT Models into ASR word graphs, emphasizing N-Best rescoring methods. The ASR hypothesis as sequence and MT hypothesis as bag of words. The CASMACAT Project [Martínez et al. (2012)]. Language understanding by generative pretraining [Radford et al. (2018)]. All these developments overcome a mile gap of development with extraordinary pace. The recent advancements in the ASR and MT model attains framework transforming the variants and multitask frameworks. For the challenge of data scarcity, and the recent development also supports data augmentation, Knowledge distillation.

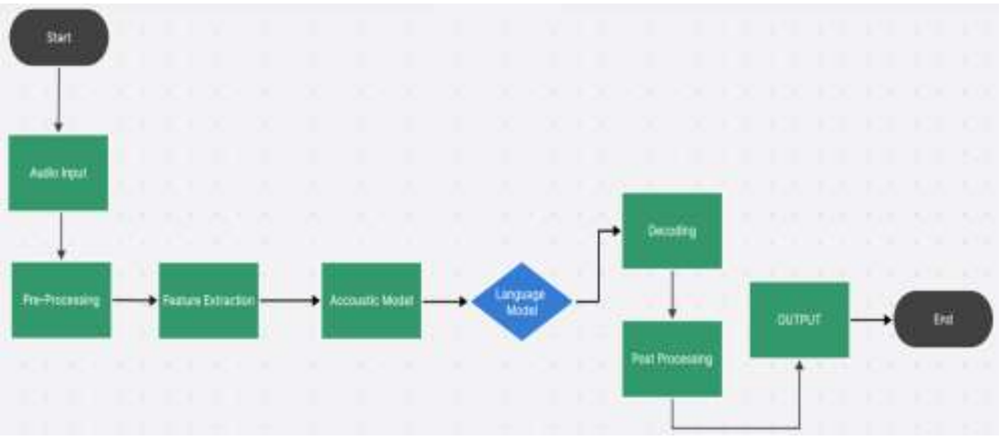
We analyse and summarize the recent advancements and limitations overcoming few gaps and still leaving scope of development into a better variant. Which include the possible integration of different ML Techniques while including user friendly and real time interaction front-end solutions.

Keywords: Automatic Speech Recognition, Machine Translation, Large language Models

INTRODUCTION

In the face of the shifting dynamics of the evolutionary techniques developed and the utilizing the specified developed model improving the capability and processing the exclusive counter efficiency the integration of Automated speech recognition (ASR) and Machine Translation (MT). To develop a more accurate and context driven Speech to Text with translation driven capability researchers have utilized multi faced approaches using different models and Machine learning Techniques. The objective of integrating the two aims to utilize the innovative exploration and utilizing the speech as input for the purpose of translation from Language A to Language B. Early solutions of the speech to text translations we can employ an ASR model to recognise the speech and translate into text using MT Model [Stentiford and Steer,1988]. Training of the model based on end-to-end speech translation provides advantages of reduced latency and more context-based modelling [Bentivogli et al., 2021], and applying the same on the dense unwritten language [B´erard et al., 2016].

The most advanced and recent approach of using the code-switching (CS) and Whisper (Whisper-large-v3) using the evaluation datasets such as ASCEND (Mandarin Specific), NTUML2021^2 (Speech Corpus) [chi-kai yang et al. (2023)].



This thesis embarks the scope of evolution of usage of two different machine learning models and integrating it. The optimized and enhanced version working and best for the present working technology. Early solutions for speech translation and Machine translation used the functioning in a way that breaks the task down into sub-groups to make it manageable sub categorical task. Similar, approaches with different path were being used. With the advancement of the data processing the scope of capability of Data usage widens and enlarges and allows the usage of LLM models for the processing of the data in the ASR model. The use LLM model has revolutionized the natural processing language and AI research. The LLM breakthrough came with the introduction of the transformer architecture in the work “Attention is All You Need” by [Vaswani et al. (2017)]. The transformer model based on the self-attention model enabled parallelization and efficient handling of long-range dependencies. The great model of generative AI like ChatGPT by OpenAI have its roots set-up using LLM. Evolution of LLM can further evolute the model ASR and integrating with Machine Translation gives the latest and the smoothest speech to Text translation with linguistic options available with wide range of the language options to translated into. LLM are a type of AI model that can process and generate the natural language text. These models are trained on massive amount of data and usage of deep learning models to identify the patterns and structures of the language. The history of the LLM can be traced back to the early developments of Natural processing Language. A remarkable review and advancement of the LLM can be significantly seen in the “Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects” by [Hadi et al. (2023)].

LITERATURE REVIEW

Human communication encompasses various modalities such as speech, gestures, and expressions, with speech standing out as the primary mode facilitating interaction among individuals. It holds significant importance due to its widespread use and effectiveness in conveying meaning. The ability to harness speech for communication with machines has emerged as a pivotal branch in human-computer interaction, revolutionizing the way users engage with technology. Automatic Speech Recognition (ASR) stands at the forefront of this evolution, contributing significantly to the advancement of artificial intelligence by enabling seamless interaction and information exchange without conventional input/output methods like keyboards. This innovation has found application in diverse domains, including assisting individuals with disabilities, enhancing automotive systems, and enabling hands-free operation in various scenarios. However, ASR's efficacy in automatic speech translation remains sensitive to recognition errors, particularly in multilingual contexts where accurate translation is paramount. Despite these challenges, recent developments in machine learning, particularly the transition from traditional models like hidden Markov models to neural network-based approaches, signify a paradigm shift in ASR research. End-to-end (E2E) modeling has emerged as a promising avenue, offering state-of-the-art accuracy in ASR tasks and paving the way for commercial deployment. This review aims to explore the evolution of ASR technologies, analyze current trends, and identify key challenges and opportunities in the pursuit of more robust and efficient speech recognition systems. Additionally, the paper examines the evolution of machine translation (MT) techniques, from early rule-based and statistical approaches to the transformative impact of Neural Machine Translation (NMT). NMT, characterized by its use of deep neural networks, represents a significant departure from traditional methods, offering improved translation quality and efficiency. Through a comprehensive analysis, this review seeks to provide insights into the advancements, limitations, and future directions of ASR and MT technologies, highlighting their pivotal role in facilitating communication and bridging linguistic barriers in our increasingly interconnected world.

The early approach to machine translation relies heavily on hand-crafted translation rules and linguistic knowledge. As natural languages are inherently complex, it is difficult to cover all language irregularities with manual translation rules. With the availability of large-scale parallel corpora, data-driven approaches that learn linguistic information from data have gained increasing attention.

Unlike rule-based machine translation, Statistical Machine Translation (SMT) [Brown et al., 1990; Koehn et al., 2003] learns latent structures such as word alignments or phrases directly from parallel corpora. Incapable of modeling long-distance dependencies between words, the translation quality of SMT is far from satisfactory. With the breakthrough of deep learning, Neural Machine Translation (NMT) [Kalchbrenner and Blunsom, 2013; Cho et al., 2014a; Sutskever et al., 2014; Bahdanau et al., 2015] has emerged as a new paradigm and quickly replaced SMT as the mainstream approach to MT.

Neural machine translation is a radical departure from previous machine translation approaches. On the one hand, NMT employs continuous representations instead of discrete symbolic representations in SMT. On the other hand, NMT uses a single large neural network to model the entire translation process, freeing the need for excessive feature engineering. The training of NMT is end-to-end as opposed to separately tuned components in SMT. Besides its simplicity, NMT has achieved state-of-the-art performance on various language pairs [Junczys-Dowmunt et al., 2016]. In practice, NMT also becomes the key technology behind many commercial MT systems [Wu et al., 2016; Hassan et al., 2018].

Objectives:

1. To understand the current working of ASR and MT models.
2. To Discuss the Advantages and Disadvantages of Current Working Models.
3. Transition from hybrid models to End – End Modelling.
4. Understanding the evolution of the Machine translation.
5. Utilizing the Role Large language Models (LLMs).

RESEARCH AND DISCUSSION

Automated Speech recognition

Automatic speech recognition is one of the most capable technologies used to detect speech using Machine Learning. Making the machines capable to understand speech and identify Languages other than Binary. Allowing the machine understand user speech with linguistic differences. It uses the approach of the Understanding the speech and converting it into series of words through a computer program, enabling a connection between human and machines.

It is also called as speech recognition with Natural Processing Language enables a gateway for the same. It can be also stated as Graphical representation of frequencies emitted at a specific instance of time.

Speech processing techniques, including speech synthesis, processing, speaker identification, and speaker verification, play a pivotal role in enabling the creation of voice interfaces, also known as Human-Machine Interfaces (HMI), and facilitating voice interaction. These techniques are instrumental in various applications across different domains:

1. **Voice Services:** Speech recognition technology facilitates the development of voice services such as speaking clocks, weather updates, and real-time race results. These services offer convenient and efficient ways for users to access information using voice commands.
2. **Quality Control and Data Entry:** In industries requiring stringent quality control measures, speech recognition technology aids in automating data entry processes. By accurately transcribing spoken input, it helps streamline operations and minimize errors in data entry tasks.
3. **Avionics and Training:** Speech processing techniques find application in avionics systems for aircraft communication and navigation. Additionally, they are utilized in training simulations, where voice-based interaction enhances the realism and effectiveness of training programs.
4. **Disabled Assistance and Vocal Dictation:** For individuals with disabilities, speech recognition technology serves as a vital tool for accessing digital resources and performing tasks independently. It enables vocal dictation, allowing users to dictate text or commands verbally, which are then transcribed or executed by the system.
5. **Embedded Voice Recognition Modules:** Voice recognition modules integrated into devices such as mobile phones and cars offer a hands-free and intuitive user experience. In cars, these modules enable functionalities such as controlling the car radio, adjusting air conditioning settings, and accessing on-board navigation systems using voice commands.

Overall, speech processing techniques have permeated various aspects of daily life, offering convenience, efficiency, and accessibility in human-computer interaction. Whether it's providing information through voice services, enhancing productivity in data entry tasks, supporting training simulations, aiding individuals with disabilities, or enabling hands-free operations in devices, the applications of speech recognition technology continue to expand, driving innovation and improving user experiences across diverse domains by [Benkerzaz et al. (2019)].

Machine Translation

Machine translation involves the utilization of automated translators to convert text from one language to another. This approach offers several advantages, including practicality, speed, and cost-effectiveness. Unlike human translator services, which may not always be readily available, machine translation provides immediate results, making it a convenient option for users. Additionally, the low cost associated with machine translation makes it accessible to a broader audience. Consequently, many individuals and organizations rely on machine translation as a necessary tool for their communication needs. The widespread adoption of machine translation underscores its importance and utility in today's globalized world. In the same way the translation of the language enables minimize the linguistic differences between Human by [Trujillo et al. (2012)]. And further Translation is the transposition of the source language text into target-language text [Poibeau et al. (2017)].

Machine translation, a field within the realm of natural language processing (NLP), revolutionizes the way information is communicated across different languages by employing automated systems to translate text from one language to another. This technological advancement has gained immense significance in an increasingly interconnected and globalized world, where the ability to overcome language barriers is crucial for communication, collaboration, and knowledge dissemination. Machine translation offers a multitude of benefits, including efficiency, scalability, and accessibility, making it indispensable in various domains such as business, diplomacy, education, and information technology.

At the core of machine translation are sophisticated algorithms and technologies that enable the seamless conversion of text between languages. These technologies encompass a diverse range of approaches, including statistical methods, rule-based systems, and neural network models. Statistical methods leverage large datasets to identify patterns and probabilities of word or phrase translations, while rule-based systems rely on predefined linguistic rules and dictionaries to generate translations. In recent years, neural network models, particularly deep learning architectures such as recurrent neural networks (RNNs) and transformers, have emerged as state-of-the-art approaches in machine translation, offering improved accuracy and fluency.

“A review on Machine translation Tools” by [Fitria (2021)]. Opens the wide thought of present idea of machine translation in the form of Google translate. Similarly the advanced idea follows everyday evolution continue and in search of better solution we process the machine translation with the speech recognition minimizing the gap between man and machine in this review paper.

The development of Automatic Speech Recognition (ASR) and Machine Translation (MT) machines has relied on a diverse array of technologies over the years. In the early stages of ASR and MT research, rule-based systems dominated the landscape. These systems utilized handcrafted linguistic rules and dictionaries to recognize speech patterns and translate text between languages. However, as computing power increased and data became more abundant, statistical methods gained prominence in the field. Statistical approaches, such as Hidden Markov Models (HMMs) for ASR and Phrase-Based Statistical Machine Translation (SMT) for MT, leveraged large datasets to identify patterns and probabilities in speech and language. Despite their effectiveness, statistical methods often struggled with capturing long-range dependencies and nuances in language. In recent years, the advent of deep learning has revolutionized ASR and MT. Neural network architectures, particularly recurrent neural networks (RNNs) and transformers, have demonstrated remarkable performance in speech recognition and translation tasks. These deep learning models learn intricate patterns and representations directly from data, enabling them to achieve state-of-the-art accuracy and fluency in ASR and MT applications. Overall, the evolution of technologies in ASR and MT has been characterized by a transition from rule-based and statistical methods to advanced neural network architectures, paving the way for more accurate and natural language processing capabilities.

RELATED WORK

Relating the research and understanding the scope of variations providing wide area of development of new ideas, provided the base for development and trying of Hybrid ASR models that uses transformers and usage of the trained LLM to extract the features from the input Speech provided, It involves the process of encoding the audio into a suitable format for the input to the LLM. Further Conversion takes place of the extracted features that can be processed by the Model.

Model Architecture to be used for preparing the Hybrid Deep learning Architecture, In HDLA we primarily focus on the formation of an acoustic model, this model utilizes various Deep learning architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Time-Delay Neural Networks (TDNNs) to process the raw audio features and extract the acoustic representations. In the same process the LLM acts as the language modelling component of the ASR model. Which also enhance the accuracy of the model.

With the preparation of the specification of the Integrated ASR and MT with the enhanced usage of LLM . The Data preparation to prepare a specific LLM model creation with the specification to enhance the capability of the ASR model while using LLM model in it. To prepare the LLM model we need to understand the requirement of access to substantial computation resources. Hence, the computation demands a tenure of time and resources, with the same the understanding and sourcing the desired data using machine learning and natural language processing.

CONCLUSION AND SUGGESTION

The ASR and MT are the two advanced working technology of NLP n Machine learning. With the level of current upsizing and the fast advancement in both is revolutionary but integrating the two with the usage of LLM can change the dynamics of the usage of both. Providing wide range of user-friendly capabilities and user-friendly interaction between human and machine.

Also increasing the accuracy of the ASR to improve the understanding capability of the ASR model resulting in the better accuracy of the Machine Translation, let it Mono-Lingual or Multi-Lingual depends upon the MT model. With the same approach there is a wider view in the research for the proposed model making it more flawless and improvising with the improving technology with better development and realistic models updated with State of the Art [SOTA] models.

REFERENCES

[1]. Nur Fitria, Tira. (2021). A Review of Machine Translation Tools: The Translation’s Ability. Language Circle: Journal of Language and Literature. 16. 162-176. 10.15294/lc.v16i1.30961.

[2]. Anugu, A., & Ramesh, G. (2020). A Survey on Hybrid Machine Translation. E3S Web of Conferences,184, 01061.

[3]. Alharbi, Sadeen & Alrazgan, Muna & Alrashed, Alanoud & AlNomasi, Turkiah & Almojel, Raghad & Alharbi, Rimah & Alharbi, Saja & Alturki, Sahar & Alshehri, Fatimah & Almojil, Maha. (2021). Automatic Speech Recognition: Systematic Literature Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3112535.

[4]. T. Chen et al., "Large-Scale Language Model Rescoring on Long-Form Data," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096429. keywords: {Additives; Video on demand; Error analysis; Lattices; Speech recognition; Signal processing; Entropy; Large-scale language models; N-best rescoring; Fine-tuning},

[5]. Zeng, Taiyao. (2022). Deep Learning in Automatic Speech Recognition (ASR): A Review. 10.2991/978-2-494069-51-0_23.

[6]. S, Karpagavalli & Chandra, Evania. (2016). A Review on Automatic Speech Recognition Architecture and Approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition. 9. 393-404. 10.14257/ijsip.2016.9.4.34.

[7]. Reddy, Aarthi & Rose, Richard & Desilets, Alain. (2007). Integration of ASR and machine translation models in a document translation task. 3. 2457-2460. 10.21437/Interspeech.2007-646.

[8]. Udagawa, Takuma & Suzuki, Masayuki & Kurata, Gakuto & Muraoka, Masayasu & Saon, George. (2023). Multiple Representation Transfer from Large Language Models to End-to-End ASR Systems.