Chapter 1: Introduction

Language is the primal form of communication between all human beings. It constitutes of all the different sounds, gestures and symbols that human beings use to express themselves and communicate with one another. It is of high importance because it not only preserves the entire history, culture and knowledge of the human society, but it also ensures the continued survival of humanity. Through language, human beings have the ability learn from one another and collaborate effectively to achieve any set task or goal. It is an extremely influential and powerful tool and depending upon how it is used, it can shape and define reality.

Although language may be expressed in many different forms, the most popular form through which it is expressed is speech. 'Speech' is a term which is used to describe the ability to express thoughts and feelings by articulate sounds. [1] According to the Ethnologue, which is regarded as the world's most extensive language catalogue, there are 7,097 languages that are being spoken in the world today and there are six major language families of the world, which are: Afro-Asiatic, Austronesian, Indo-European, Niger-Congo, Sino-Tibetan and Trans-New Guinea. [2]

Even though communication through speech had always been reserved as a special activity that ensued only between human beings, the advent of unprecedented computing power and seemingly limitless technology, such as Artificial Intelligence, in today's world has enabled mankind to share the activity of speech with machines.

Enabling computers to understand speech is a very important task because, with the increasingly high demand that is being placed on computers for the performance of more complex tasks, there is the need to develop a way in which these complicated requests could be made to the computer system in a very easy and natural way.

Natural Language Processing (NLP) is the area of computer science research that enables computers to understand and manipulate natural language text or speech. It involves understanding how human beings learn and use natural language so that technological tools and systems can be built to utilize natural language to perform desired tasks. It began in the 1950s and has since then evolved. According to the research paper entitled “Natural Language Processing: a historical review”, Natural Language Processing has gone through 4 different phases of development since its inception. The first phase was driven by Machine Translation, the second phase was influenced by Artificial Intelligence, the third phase was the “grammatico-logical” phase and the fourth phase was heavily focused on lexical and corpus data. [3]

In the first phase, machine translation in NLP was implemented as a lookup, where each word was processed individually by checking its value in a given dictionary. This approach was problematic because it led to the formulation of syntactic and semantic errors. As a result of this, a large portion of this first phase of NLP was heavily focused on syntax processing strategies which could help mitigate the errors caused by the approach which was being used.

In the second phase, NLP was coupled Artificial Intelligence and emphasis was placed on ‘world knowledge’ and the role it played in the construction and manipulation of meaningful representations. The approach adopted in this phase of NLP involved the construction of knowledge bases to satisfy certain specific user input. The real challenges faced by NLP researchers in this phase was developing a general purpose front-end and providing for the acquisition of application-specific knowledge to handle a user real needs in dialogue.

In the third phase, NLP systems were developed based on grammatical theory. In this phase, NLP researchers developed a range of grammar types which could not only be computed but

could also be parsed. Therefore, the processing paradigm for the analysis of an input sequence was based on the interpretation of the syntax of the input sequence into a logical form.

In the fourth phase, lexicons were used to replace syntactic general rules in the development of NLP systems and there were significant advances in NLP technologies such as Speech Recognition. This phase was also characterized by a rapidly growing interest in the development and provision of linguistic resources.

NLP is a very important piece of technology because it not only provides us with a more native way of interacting with computer systems, but in the grander scheme of things, it also possesses the ability to enable us to make meaning out of immense and seemingly unrelated pools of data. Some popular, real-world applications of NLP can be found in: customer or personal virtual assistants, text summarization systems and chatbots.

Natural Language Processing systems need a lot of human language data to be able to work effectively and efficiently. Some languages have a lot of already recorded and easily accessible data which can be used to build NLP systems. However, there are many other languages that do not have a lot of recorded data and thus do not have effective and scalable NLP systems built for them. These languages are referred to as “low resource languages” and native Ghanaian languages fall in this category.

This thesis work is going to explore the various approaches to building NLP systems for native Ghanaian languages and look at the different ways in which the data for developing these systems can be obtained.

Research questions

1) What will it take to develop an NLP system for a low-resource language?
2) Is there a less costly and more efficient way of building an NLP system for a low-resource language?

Chapter 2: Literature Review

Developing NLP systems for languages that are identified as low-resourced is a very challenging task because one may have to overcome certain peculiar problems such as the lack of an already existing standard written representation of the language.

A Speech Translation System was developed for the low resource language of Pashto and it was to be applied in the domain of medical exchanges between a medical official and a patient. The prototype of this system was implemented with 2 speech recognizers, 2 parsers, 2 speech synthesizers and a user interface. Before this system could even be built, issues relating to the orthography of the language had to be resolved. Since Pashto does not have any standardized writing systems or spelling norms, acoustic data relating to the language had to be transcribed directly into a phonemic representation of the language. Another element that was needed to build the system was a corpus of the language. This corpus was obtained from the manual transcription of a series of "Voice of America" Pashto Service broadcasts. After the corpus was obtained, two test data sets were created. The first test data set consisted of 5,128 words and was created from 5 dialectically diverse speakers. The words that were in the first data set were not a close match to the medical dialogues that could be found in a large portion of the language model data. The second test data set was made up of 3,409 words and was created from 4 Eastern Afghan Pashto speakers. The words in the second test data set were a good match to the medical dialogues found in the language model data. After the creation of the test sets, a language model was built. As a result of the morphological complexity of the language and the small amount of available training data, the language model that was built had more fine-grained back-off layers than a traditional n-gram language model. In the language modelling process, a clustering tree was created using the minimum discriminative information clustering algorithm for the vocabulary. The root of the clustering tree represented the whole vocabulary while every other node of the tree represented a class that consisted of all of the words in the

descendant nodes of that particular node. Finally, to evaluate the accuracy of the speech recognition system, the Word Error Rate (WER) for a test set consisting of 272 words, from a single Eastern Afghan speaker, was used. The initial WER obtained was 21%. But however, after changing the orthographic points in the data used, the WER ranges from 19.4% to 29.7%. [4]

A proposal was made strongly suggesting Cross-Knowledge Transfer using Multilingual Deep Neural Networks with shared hidden layers. This is because, Shared Hidden Layers Multilingual Deep Neural Networks (SHL-MDNNs) can reduce word error rates by 3-5% respectively, for all languages which it can decode. In SHL-MDNNs, the hidden layers of the neural network are made available to many languages whiles the final softmax layers are made language dependent. Using the SHL-MDNN is proven to be beneficial because it not only produces a much lower word error rate when compared to Monolingual Deep Neural Networks (MDNNs), but it also provides the flexibility of being used for the benefit of other languages even if those languages are phonetically far away from the source languages used to train the SHL-MDNN. This process is known as cross-lingual model transfer (CLMT). During the procedure for CLMT, the shared hidden layers of the already existing SHL-MDNN are extracted and a new softmax layer, whose output nodes correspond to the senones in the target language, is placed in the extracted layers. For the final step in the CLMT process, the new softmax layer is then trained with data from the target language. Because of the versatility and efficiency of SHL-MDNNs, they are proposed to be the central framework in the development of a universal Automatic Speech Recognition System. [5]

An end-to-end speech-to-text translation system was built to learn or decode speech without using source language text. The implementation of the system was based on the *seq2seq* model implemented by TensorFlow and even though the system reused some components of the model provided by TensorFlow, it also provided its own additional features such as: a

bidirectional encoder, a beam search decoder, a convolutional attention model and a hierarchical encoder. The system was then trained on the French-to-English BTEC corpus using a synthetic speech generator called “Voxygen” and the Adam algorithm, with an initial learning rate of 0.001. For each input into the speech model, the raw data of the speech was segmented into frames of 40 ms, with step-size of 10 ms, and 40 MFCCs were extracted from each frame, along with the frame energy. The system was tested after it was trained, and it generated a Word Error Rate between 23% to 26%. Even though the Word Error Rates of the system are not very low, they can be mitigated because of the synthetic nature of the data used to train and test the system. [6]

REFERENCES

[1] English Oxford Living Dictionary. Speech. Retrieved from: https://en.oxforddictionaries.com/definition/speech

[2] Ethnologue. Summary by language family. Retrieved from: https://www.ethnologue.com/statistics/family

[3] Karen S. Jones. 2001. Natural Language Processing: a historical review. Retrieved October 8, 2018 from https://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf

[4] Andreas Kathol, Kristin Precado, Dimitra Vegyri, Wen Wang and Susanne Riehemann. n.d. Speech Translation for Low Resource Languages: The Case Of Pashto. Retrieved October 8, 2018 from https://pdfs.semanticscholar.org/c365/d34875fdbb62d9e810c42445cf9b49832236.pdf

[5] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng and Yifan Gong. 2013. Cross language knowledge transfer using multilingual deep network with shared hidden layers. Retrieved October 8, 2018 from https://ieeexplore.ieee.org/abstract/document/6639081/

[6] Alexandre Berard, Olivier Pietquin, Laurent Besacier and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. Retrieved October 8, 2018 from https://arxiv.org/abs/1612.01744