Project Proposal

Project Title: Using Machine Learning and a New Shorthand for Faster Transcription

Author: Ronit Avadhuta

Date: 10/12/2020

Phrases:

Phrase 1: The process of digitally transcribing information does not meet our current needs of efficiency in our fast-paced world.
Phrase 2: How can we find/develop a faster method for data input that does not need specialized equipment such as a steno and is accessible to everyone?

Project Definition: .....

The aim of this project is to quickly send text in a digital medium. I will utilize a machine learning algorithm to learn to recognize characters in digital writing from the new shorthand I developed last year. The second part of the algorithm will then convert the Digital Shorthand Key (DSK) into English text. I will then train the algorithm to recognize the characters via training and testing data and measure loss. I will also write a series of rules to turn the shorthand into English from the DSK. This program would prove to be useful for those in time sensitive situations where they do not have time to type up a message to send on a traditional keyboard. Machine learning well adapted to these tasks such as Natural Language Processing and image recognition and transcription. Furthermore, numerous qualities of shorthands have been examined in order to help make the Digital Shorthand Key that this project will be using.

Background:

Our society demands efficiency. This is evident with several recent inventions, such as with 5G and online medical treatment. However, while text messages can take less than 5 seconds, we are still typing on the same QWERTY keyboard since 1874. We are not limited by speed technology works but rather the speed at which we type with our fingers. And as part of the Information Age, I believe it's time to change that.

The process of digitally transcribing information does not meet our current needs of efficiency in our fast-paced world. How can we find/develop a faster method for data input that does not need specialized equipment such as a steno and is accessible to everyone? This is where this project comes into picture. To better understand this project, some background knowledge on shorthands and machine learning would be beneficial.

Major Sections:

Shorthands:

Shorthands are the art of transcribing words and information (in English) quickly by hand, since traditional words take too long to write. For the past few hundred years, Gregg shorthand has been widely considered one of the most popular and effective shorthands (4). Pitman shorthand is also a very popular shorthand used by many professionals to record information, but relies on the thickness of stroke (6). However, my project last year focused on the number of strokes and how many pixels they took up (all being the same length). For these reasons, Gregg seemed to be the better option. The new shorthand that could beat Gregg thanks to a smaller phonetic inventory represented by only 12 distinct characters. There are numerous benefits to a phonetic system, including future application in different languages too (3).

By reducing the number of distinct characters, it messes up Zipf's law and helps maintain security. Zipf's law is a law that denotes the exponential decreasing manner in which different letters and phonemes in the English language are distributed. This is a common method of trying to decode encrypted information but parsing information through a Digital Shorthand Key (DSK) would prevent that possibility. Not to mention, the whole system could be converted to a 7-bit version of Unicode that would save space and make the information more secure.

Machine Learning (image recognition):

Image recognition is an important part of this project because it is the first step of getting the shorthand from an image into a format where functions can be performed to get it into English. This can utilize one of two methods (or both): blocking and weighted matrices. Blocking is a method in image recognition that uses distinguishing characteristics to split the image into parts that the algorithm can tackle one by one (8). From this point, these sections of the image can be split up into rows and columns in which certain areas have more weight than others and can be represented with vectors (3).

The output of the first Machine Learning program will output information written in the DSK, however this is not English. In order to get it back into English, lexers and parsers will have to be used, or another Machine Learning algorithm that strictly deals with text to put the information into context (7). Furthermore, although the information about English words in IPA is phonetic like the shorthand, the IPA will have to go through a dichotomy in order to resemble the DSK.

Experimental Design/Research Plan Goals:

Project variables:

IDV: The iteration being used

DV: Loss, Accuracy, Speed, and other values TBD

Controls: The medium through which they will be processed and the DSK mapping

Materials: No more than a computer. Possible programs that may be utilized include: Google's Teachable Machine, Tensorflow, and Jupyter Notebooks.

Procedure:

TST (text to shorthand to text)

- Assemble English Text
- Translate to DSK
- Translate back using an iteration
- Measure accuracy
- Change model and repeat
- Use new text

WST (writing to shorthand to text)

- Assemble shorthand characters
- Parse through system
- Measure accuracy
- Assemble shorthand syllables
- Parse through system
- Measure accuracy
- Assemble shorthand sample text
- Parse through system

Risk/Safety Concerns:

This project raises no safety concerns.

Data Analysis:

Data analysis will, for the most part, be conducted by the programs that I choose to use. Nevertheless, large data sets such as those pertaining to image recognition can be recorded in Excel whereas loss and other quick metrics can be noted in my logbook (bullets or tables).

Potential Roadblocks:

Image Recognition:

Recognizing which index the written DSK characters connect to is a job that will need to be done using some sort of Machine Learning algorithm. There are many options but I will probably end up using something similar to Jupyter Notebooks because I will need to shift through lots of data. Although Google's teachable machine also appeared to be effective in recognizing each piece, when the characters are combined into phonetic blocks, it fails.

DSK Recognition:

Recognizing which words in the shorthand relate to English may raise some difficult problems because it is not a one-to-one translation. My current plan is to utilize a pre-existing Excel document which matches English words to their IPA to help with this, but if this fails tagging might be necessary.

Homonyms/tagging:

I thought of the concept of tagging in order to successfully distinguish words with the same phonetic reading yet different transliterations. This process would give no tag to the most common (and hopefully) only variant while the next ones receive higher numbers. This process may have to be used for certain proper nouns too.

References:

(3) Kotipalli, Krishna V. Phonetic-based text input method. United States US8200475B2, filed February 13, 2004, and issued June 12, 2012. https://patents.google.com/patent/US8200475B2/en.

(4) R. Rajasekaran, Dr. K. R. (2014). Statistical review of Online/Offline Gregg Shorthand Recognition using CANN and BP - A Comparative analysis. *IJAIST*, *24*(24), 13.

(6) *Automatic recognition and transcription of Pitman's handwritten shorthand—An approach to shortforms—ScienceDirect*. (n.d.). Retrieved October 15, 2020, from https://www.sciencedirect.com/science/article/abs/pii/0031320387900082

(7) *US7027974B1—Ontology-based parser for natural language processing—Google Patents*. (n.d.). Retrieved October 15, 2020, from https://patents.google.com/patent/US7027974B1/en

(8) Mohamed, A., & Rohm-Ensing, E. (n.d.). *English-Arabic Handwritten Character Recognition using Convolutional Neural Networks*. 7.

**

Timeline: