TAP Report

Michael Donovan and Thomas B. Kinsman
National Technical Institute for the Deaf (NTID)
Rochester Institute of Technology (RIT)

Spring 2023

Abstract

To provide fair and equal access to a college education for deaf and hard of hearing students, college lectures given by speaking professors must be translated into another language. Spoken English must be converted into either written English or ASL (American Sign Language). This is not a simple speech-to-text process, but a complex translation problem which requires situational awareness and context on the part of the translator. The people who do the translation (transcriptionists and interpreters) must perform the translation in near real time to promote inclusivity in the classroom. There must be only a short time delay between the speaker and the resulting translation. The language conversion must provide a form of communication that is accurate and complete. The task requires a human-in-the-loop to make the language conversion correctly, and capture nuances of the language.

RIT already has a valuable confidential source of information that they can TAP into: previous transcriptions. This project demonstrates that the analysis of prior transcripts can prepare the captionists and interpreters for uncommon words, unusual words, and difficult words that will occur in a lecture. This resulting software project can prepare students and staff, for words that will be included in the coming lectures, so that no one is suddenly surprised by unfamiliar terminology. An pilot project demonstrated that this software and this technique results in smoother transcription, more accurate ASL, and an overall better experience for everyone involved in the lecture (including hearing students).

1 Overview

TODO - this needs revision once the rest is written.

2 The Transcription Challenge

A direct, speech-to-text conversion of spoken English introduces problems that complicate understanding for students, and can make things worse for them instead of simplifying the understanding. Presently, fully automated systems inject errors in two ways. First, fully automated systems include words which should have been removed. Secondly, fully automated systems convert new and unfamiliar terminology to the wrong words.

2.1 Words Which Should Be Removed :

A straight, direct, speech-to-text program copies all of the spoken disfluencies. Disfluencies are spoken hesitations that are inserted in spoken English to indicate that the speaker still wants control of the conversation, but is considering and composing the next section of the conversation. These are words such as: "like…", "um…", "ah…", "err…", and "so…". All of these spoken utterances indicate the speaker is still working on what to say next, and wishes to retain control of the conversation flow, yet are not needed in the transcription process.

2.2 Words Which Must Not Be Removed:

Verbatim, conversion from audio to written transcription leaves in these hesitations and filler words which we would prefer to have edited out. Automated speech-to-text also have issues when trying to decide which stop words should be removed from the translations. For example, the word "of" is often ignored as an unimportant word in natural language processing. Yet, in the context of a lecture the word "of" is important. Phrases like, "center of mass", "King of Spain", and "basis of comparison" require that the word "of" be retained in the text.

2.3 Inaccurate Translations are Worse than Missed Translations:

With the recent publicity of large language AI models, comes a misperception that AI can do almost any task desired. The statistical AI models do not replace human thought. People must be used to make sense of and validate the work. The process of converting spoken lectures into written transcripts of English is not fully automatic, yet.

To prove that computer models are inaccurate, consider the problem of texting using a cell phone. In the context of typing a text, even the letters physically typed by a human are not correctly converted into English. One example found was the sentence, "Mom, this is Janet, I am coming home for the weekend, and I am bringing drugs." The words uttered were "Mom, this is Janet, I am coming home for the weekend, and I am bringing [my boyfriend] Doug."

What the computer heard was, "Mom, this is Janet, I am coming home for the weekend, and I am bringing dug." However, the computer could not make sense of what it thought the speaker said. The computer tried to make sense of the expression "... bringing dug", which makes no logical sense. To compensate, the computer put in words that it thought made betters sense, and substituted in the text, "... I am bringing drugs."

In fact, even the expression "speech to text" is not reliably converted correctly. A speech to text system often converts the expression "speed to text" into the words "speech detects."

2.4 Analogy of Reading Handwriting:

Perfectly accurate transcription of speech will always require a human-in-the-loop to correctly translate the spoken words into written their English equivalents. Transcribing is analogous to converting written script into typed English. For example, consider figure (TODO reference, Following figure). In this process a human has to read the words, correct mistakes that happened along the way, and then type the desired output.

While computers are making great strides at understanding human speech, they are not perfect. Even if they are as good as people, they would still confuse homophones (would that sound alike). The sound for 'B" could be a single letter of the alphabet, an insect that pollinates plants, or a verb.

By analogy, handwriting recognition is not well understood in a generic sense yet. The best handwriting analysis is per-person. Nevertheless, after 30 years of marriage, Dr. Kinsman cannot read Mrs. Kinsman's handwriting. When he goes grocery shopping with a grocery list written in script, he has to skip purchasing items for which the script is unintelligible.

Given this scenario, one would naturally ask, why doesn't Mrs. Kinsman hand print the shopping lists? After all, OCR is pretty good now. To illustrate why this is a problem, consider the following example. Dr. Kinsman has been taking notes using printing for 40 years. His printing is very consistent. He would like to be able to train an OCR system to convert his scanned notes into ASCII text. To make the problem simple, he always prints in all upper-case letters.

In the (TODO Following figure) a sample of all upper-case letters was fed into an OCR engine. While humans can clearly read it, the results also clearly demonstrate that a straight computer conversion is not correct.

3 Ethical Barriers:

OCR on the above figure generates the following output:

1. I ALWAYS UPPER CASE, "2" PRINT LETTERS BUT THE AND "C" LOOK ALMOST SAME AT WHEN I FULL CANNOT CORRECTLY SPEED READ MY EITHER. TAKE LETTER THE NOTES A COMPUTER TYPED LETTERS

4 Conclusions

• This processing isolates words that can be used to prime, or prepare, transcriptionists, captionists and interpreters. • By preparing the transcriptionist, we can drive down the error rate. AND in some cases, transcriptionists or interpreters skip entire sentences because they are need to look up complicated words. • In addition to showing the word summaries could be produced, we also discovered some natural

5 TBD - BELOW THE LINE

Speech to Text does not filter out words correctly.

”Damn autocorrect” – machine methods for fixing spelling error and grammar errors actually inject mistakes. Many

Complexity of the problem: - Ethical Issues: - Ethics issues: using the voice of professors and students without explicit consent. - This is why we focused on Kinsman’s lecture. (Ignore other ethical issue.)

GOING TO AI: We are not isolated. We are aware that there are AI language models working on the transcription problem as well. The results of this TAP analysis can be used to help train future AI models. It would be a natural fit to integrate this analysys with future work on AI models for transcription.

Regardless, AI does not replace the need for ASL translators.

Scope of the problem.

Experimental 1 Description: - Mike - Summary from Mike. - The traditional methods do not apply in this case. - Typical ”noise” words are actually import to us. = Example some ”stop words” are not stop words - ”Basis Of Comparison” – search the text for concordances for ”Comparison” to ”Basis”. Example: Project Project for the course, versus Project (as in projection vector) Homonyms – pronounced differently, but spelled the same.

- ”Center Of Mass” - ”King of Spain”

Results for Experiment 1 - TFIDF - Finds the words to do.

What was done? - Mike

Results - Mike

Discussion - Joint Future Work - Can be used to prime an AI system. . .

Conclusion

PRIMING THE CAPTIONISTS HELP

Other Experiments

Psychological barriers

Credits

Thanking fast and slow

5.1 SS-QQQ

6 Section-QQQ

7 Key Take-Aways

$$Distance = S = V \times \Delta t \tag{1}$$

$$I(x, t = 0) = I(x, t = 0) = I_A \tag{2}$$

$$I(x, t + \Delta t) = I_B \tag{3}$$

$$I_B = I_A - \frac{\Delta I}{\Delta t} \times \Delta t \tag{4}$$

8 Conclusions