Self-correction of automatic speech recognition

Zihao Lin

Department of computer science, University of Wisconsin-Madison, Madison, 53706, United States

zlin329@wisc.edu

Abstract. The rapid development of modern technology and society has raised the need for a large number of real-time translations, such as Automatic Speech Recognition (ASR) technology. One of the important factors determining the fluency of cooperation is that ASR translates the input speech into readable text so that both parties can understand each other's meaning. There is also a huge demand for and reliance on automatic speech recognition technology for modern applications. This article focuses on the word matching algorithm for backward search (WMA-BS), which requires two separate steps, one for the word matching algorithm and the other for the backward search step, for the first step, which can also be done by applying the word-level N-Gram language model. n-Gram model is a language model in which when n words appear in a text, we can use these n words to predict the structure of the text. we can use these n words to predict the structure of the text. In order to make the word-level N-gram language model suitable for the method, it is more useful for us to integrate it into shallow fusion. One of the most important parts to be addressed in the word matching algorithm is the boundary, or distance, between two words in English.

Keywords: Automatic speech recognition, real-time translation, text to word.

1. Introduction

With the rapid development of modern technology and society, more and more data transfer and international cooperation are need. This raises plenty demand of real-time translator such as Automatic speech recognition (ASR) technology [1]. One of the important factors to decide if the cooperation is fluently the preciousness of the Automatic speech recognition translate the input voice to a readable text in order for both side of understand other side meaning. Applications in the modern society also have huge demand and rely on the ASR technology, for example, Wechat in China, needs to deal billons and billons inputs voice and translate to an readable text for user for such an simply "translate to text from voice" button, and no so much input voice is easy to translate and re-written as an readable text because it many accompany by background noise and some difficult to understanding accent, and some input voice even have two language mix in one sentences to make the machine difficult to analysis what the user said and translate to a readable text [2].

From the educational area, the preciousness of ASR technology is also essential in this time. Due to the COVID-19 virus, many Universities or schools now need to offer online class for student either for synchronous or asynchronous class. Student who studies in different country might need translate for their first year, such as zoom now offer a real-time translator. And even for the in-person lecture, there are also exist demand for real-time translator for those students who need more time need to understand the lecture [3-5]. So, the low percentage preciousness of ASR will cause a barrier for the business develop and delay the efficient of the cooperation, the convenience of communication, and also the waster unnecessary time for study.

Therefore, to deal with those problems or even make some progress that help the ASR becomes more precious and has less grammatical errors in the translated re-written text, we need to formular some model or give some strategy for it, which inspired by the Word-Matching Algorithm with Backward Search and improving the readability for Automatic Speech Recognition Transcription. In those two methods, it used End to End model as a major method to correct the grammatical error in the ASR rewritten text in order to make it make more sense and becomes understandable, with accompany by method such as CLAS (contextual listen, Attend, and Spell) and OOV (out-of-vocabulary) to improve the correctness of the word-matching algorithm with backward searching. And also, to deal with the readability of the ASR rewritten text, it utilizes the method such as sequence-to-sequence (seq2seq) generative model to handle the proposed task, Natural Language Processing, Pre-trained Language Model, Dataset Construction to make the rewritten script become more readable for user.

2. Language Module

To help self-correct the output script of Automatic speech recognition (ASR), there are several ways for us to make the translation more precious [6-8]. One of the methods is apply language module for end-to-end ASR, which here we are using n-gram Language module. The other one method is using the Word-Matching Algorithm with Backward Search algorithm. Both methods are using for self-correcting the transcript of the output of ASR if there is any word or sentence that obvious translated not preciously or having some grammatical error, they will help correct those errors. To help integrate the language module in the end-to-end ASR, one of the widely used approach is shallow fusion. Shallow fusion refers to log-linear interpolation with a separately trained language model at each step of the beam search.

2.1. Word-Matching algorithm with backward search

One of the methods to correct the output of the ASR is Word-Matching Algorithm with Backward Search, which is a method that based on the CLAS. CLAS is based on the LAS (Listen, Attend, and Spell) by it will adding some bias code in the translated test to help machine analyse that which word or sentence need to correct and then it will give a </bias> label to help machine do the further operations. With the </bias> labels, the machine can use the word-matching algorithm with backward search (WMA-BS) around the target ward and then it can compare the target words with all contextual phrase with similar meaning or pronunciation in order to determine if the target words around the label need to replace with the phrase or not to make the script more reasonable and understandable [9, 10].

To accomplish the word-matching algorithm with backward search (WMA-BS), two steps need to accomplish separately, one is the word-matching algorithm, and the other one is the backward search steps, For the first steps, it can also accomplish by applying the Word-Level N-Gram language models The N-gram model is a language model that when there are n words appear in the text, we can use those n-word to predict the structure of the text. To make the Word-Level N-Gram language models fit the method, it is more useful for us to integrate it into shallow fusion. One of the most important parts to solve in the word-matching algorithm is the boundary between two words in English, or the distance. For instance, ASR may have difficult time to figure out is the human saying “hometown” or “hometown”, and this can be cope by score the distance for the phase so that it can find the minimum distance of the phase. For the language that is difficult find the boundary between words, such as Chinese and Japanese, it can solve by “converts character-level sequence into a word-level sequence”. However, to improve the readability of the output script, we can construct dataset for the ASR for it to compare the phares in the output with the dataset and to correct its error, such as grammatical and translated errors. As well as English, we can also use dataset to help we address the translation of different language problem, such as language like Chinese and Japanese, which doesn’t have a very clearly boundary between wards. In specific, we can build the dataset base on what we know so far about the vocabulary and use it to convert character-level sequence into a word-level sequence.

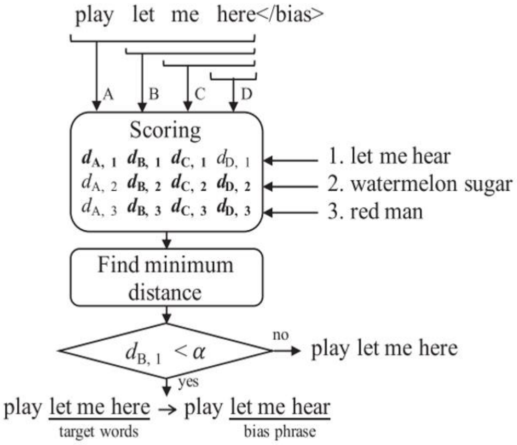


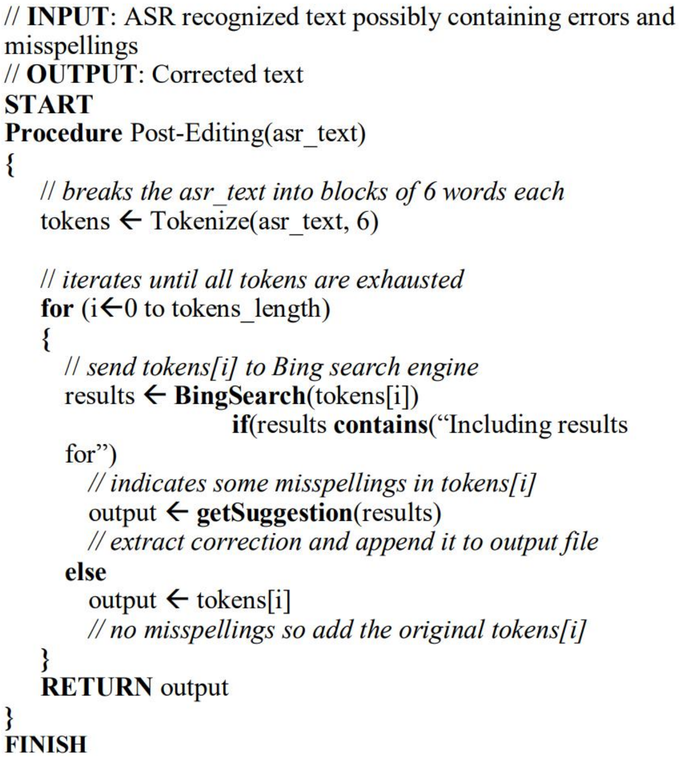
Figure 1. WMA block diagram [3].

2.2. Dataset construction and inspiration

To make the output of the transcript become more readable, we can construct dataset for us to checking each token of the sentences or ward in order to make it more preciousness when translating. To fully constructing dataset and use it to help us correct the grammatical and translated errors, we can use a text-to-speech (TTS) system help us translate our original sentences to the sentence that having grammatical and translated errors. Since we need to consider the conversation between daily situation, we also need to know that daily conversation often contains much grammatical error. Therefore, we can first simulate the error that human might have during different conversation or situation, then we apply those sentences into TTS systems and transcribing to ASR system, in this way, we can simulate different both grammatical errors and ASR errors. In addition, it is difficult to take every situation and every kind of grammatical error under consideration, we can build different dataset for the translation system, for which people can using different dataset under different situation. For example, people who need to having commercial cooperation between different counties can using the commercial dataset which we might build it before and including the languages that they need to translate.

2.3. Backward Search

Another important stage for implement Word-Matching Algorithm with Backward Search for self-correcting is the backward search stage. Since the CLAS cannot decoding the label </bias> in hundred percentage correct, it needs to recalculate the distance of the words to make the decoding of the label more preciously. After the backward search done, the scripts are roughly corrected, however, we can further correct the output scripts by the post-editing error and misspelling. One of the methods that can help to check the errors and misspelling is that using the Bing's spelling suggestion to help check the text and fix, which the Bing's spelling suggestion using the error correction algorithm to help detect the errors and doing the post error correcting.



```
// INPUT: ASR recognized text possibly containing errors and
misspellings
// OUTPUT: Corrected text
START
Procedure Post-Editing(asr_text)
{
    // breaks the asr_text into blocks of 6 words each
    tokens ← Tokenize(asr_text, 6)

    // iterates until all tokens are exhausted
    for (i←0 to tokens_length)
    {
      // send tokens[i] to Bing search engine
      results ← BingSearch(tokens[i])
                        if(results contains(“Including results
      for”)
        // indicates some misspellings in tokens[i]
        output ← getSuggestion(results)
        // extract correction and append it to output file
      else
        output ← tokens[i]
        // no misspellings so add the original tokens[i]
    }
    RETURN output
}
FINISH
```

Figure 2. Procedure of the ASR [11].

3. Conclusion

The rapid development of modern technology and society has raised the need for a large number of real-time translations, such as automatic speech recognition (ASR) technology. One of the important factors determining the fluency of cooperation is that ASR translates the input speech into readable text so that both parties can understand each other's meaning. There is also a huge demand for and reliance on automatic speech recognition technology in modern applications. By the large amount of algorithm, models, and dataset, we can use that to predict how the translator will work by those factors. The specific work of translation is taking the input argument, which is the speech of human, translate to a readable script. As the voice input been receiving, the algorithm will separate the whole sentence into different tokens, which is character sequence. We can apply the word matching algorithm to convert the character sequence into a word. After we get each of the word and sentence, it then applies the backward search to help correct the spelling and the error, which can by the Bing’s suggestion. On the other hand, it can compare the output by the dataset which build before in order to correct the error, both ways can help to make the scripts more readable and precious close to the input.

References

[1] Pantel, P., and Lin, D 1998 *Madison Wisconsin*, 95-98.
[2] Yu, D., and Deng, L 2016 Automatic speech recognition **1**.
[3] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... and Wellekens, C. 2007 *Speech communication*, **49**, 763-786.
[4] Juang, B. H., and Rabiner, L. R 2005 *Georgia Institute of Technology*. Atlanta Rutgers University and the University of California, Santa Barbara, **1**, 67.
[5] Ghai, W., and Singh, N 2012 Literature review on automatic speech recognition. *International Journal of Computer Applications*, 41(8).
[6] Al-Shabi, M., Shak, K., and Tan, M. 2022. ProCAN: Progressive growing channel attentive non- local network for lung nodule classification. *Pattern Recognition*, 122, 108309.

[7] Potamianos, G., Neti, C., Luettin, J., and Matthews, I 2004 Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22, 23.

[8] Zhu, W., Liu, C., Fan, W., and Xie, X. 2018. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In 2*018 IEEE winter conference on applications of computer vision (WACV)* 673-681.

[9] Cooke, M., Barker, J., Cunningham, S., and Shao, X 2006 An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, **120**, 2421-2424.

[10] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. 2019. Self-attention generative adversarial networks. In International conference on machine learning, 7354-7363.

[11] Bassil, Y., and Alwani, M. 2012. Post-Editing Error Correction Algorithm for Speech Recognition using Bing Spelling Suggestion. *International Journal of Advanced Computer Science and Applications*, **3**, 2.