Improving Speech-to-Text Transcriptions of Chinese Podcasts

Elliot Schmitt
Chinese Flagship Program
Indiana University
Bloomington, IN.
jawschmi@iu.edu

Abstract

The internship was spent contributing to the Tech Center's ongoing podcast project; an application that will collect language podcasts and extract information from those podcasts that can help language learners and instructors better find relevant language learning materials. The podcast audio files are transcribed by software, and most of the work of the internship was creating a markup tool that can improve the quality of the podcast transcriptions. The transcriptions were corrected by hand and then a rule-based approach was developed to correct errors the transcription software consistently made. This adds a layer of polish to the project, yielding cleaner and more accurate English translations later on. The internship was largely exploratory, and the rest of the time was spent experimenting with other aspects of the project, such as researching lexical sophistication and how a metric for the sophistication of a text could be useful information to teachers or learners trying to gather useful study materials.

1 Introduction

The Language Flagship Technology Innovation Center is developing an application to collect language podcasts, and to extract information from those podcasts, such as transcriptions, translations, summaries, and a list of named entities. The purpose of this application is to help language learners and instructors find relevant language learning materials very quickly. The application will be multilingual, and so far information from Russian podcasts has been collected and extracted. This internship was spent helping expand the project to Chinese.

2 Markup Tool

This project has an order of operations, with transcription of the podcast audio preceding translation of the transcripts into English, which precedes summarization and the extraction of information. The bulk of the internship was spent creating a markup tool that polishes Chinese transcriptions before they are translated. The podcasts used come from Radio France Internationale Chinese and all relate to environmentalism. The transcription software is Google's Speech-to-Text API, which is the state of the art.

To create the markup tool, it was first necessary to understand what kind of errors Google's API was making. If the program makes similar errors in a consistent way, then it should be possible to write software that can fix the error. After listening to the podcasts and marking up errors by hand, it was clear many of the errors relate to interference in the audio file (e.g. when music was played while the podcaster was speaking). As acoustic issues are difficult to prevent and unpredictable, it is important to note that not all errors can be removed. But many errors are much more manageable, and can be grouped into a few categories: homophone errors, ordinal numbering, "的sentence breaks," and percentages.

Homophone errors are when the characters in the transcript had similar pronunciation but a different meaning than what the speaker had said. This produced nonsense translations. Consider the table of examples on the next page.

These examples are not all pure homophones, they may differ slightly in tone or pronunciation, but they do sound similar. preventing this is a matter of improving speech-to-text software, an ongoing field of research. To repair this kind of error, a machine-learning approach (discussed more below) was attempted, but ultimately proved unsuccessful. The markup tool includes a dictionary of several common homophone errors and have the program simply replaces the wrong words with correct ones. This was easy to code, but only repairs the ho-

区周, 欧洲	一情, 疫情	牙吗? 写, 亚马逊	80%期, 87%
Ou1zhou1	yi1qing2, yi4qing2	ya2 ma xie3, Ya4ma3xun4	bai3fen1zhi1qi1
Ou (surname) week, Europe	one feeling, epidemic	tooth ? write, Amazon	80% cycle, 87%

Table 1: Examples of homophone errors found in transcripts. Characters left of the comma are the errors, the right of the comma are the characters the podcaster actually read

mophones that were noticed and homophones that were clearly meaningless (to avoid replacing characters that are correct without realizing it). It improves the transcriptions, but only slightly, and only for those specific terms.

The ordinal numbering issue relates to the characters in Chinese 的and 第. 的(pronounced as de or di) is the most common character in written Chinese and has many different uses. 第(di4) is also a common character, but its usage is much more narrow; being used for ordinal numbering (i.e. marking items as first, second, third, etc). It was very common for the transcription software to write 的instead of 第, possibly think because 的can also be pronounced as "di." For example, the transcriber yielded

格拉斯哥的26家气候峰会期间
Glasgow of 26 climate summit during
"During Glasgow's 26 climate summits"

Where the 的should instead be a 第that marks the event as the 26th climate summit. The markup tool fixes this error most of the time by making rules to identify when a 的is placed where a 第should be and replacing the wrong character with the correct one. This makes the correct translation "During Glasgow's 26th climate summit."

"的sentence breaks" is another issue relating to the Chinese 的character. 的is most commonly used as a possessive, similar to the 's in English, or to modify nouns with adj + 的+ noun. 的can also be part of a grammatical construction where it is placed at the end of a sentence for emphasis. For example,

他们是在网上认识的。
They are at online met
"They met online."

The 是(shi4, meaning to be) is often omitted. Consider this excerpt produced by the transcriber: "来保护环境和居民的。健康因为新冠疫情..." which translates as "To protect the environment and residents. Health because COVID-19..." There is an instances of a period following a 的when the 的in that sentence serves a different function. In this example, the 的is supposed to join the two nouns 居民(resident) and 健康(health), but the misplaced "。"causes what should be part of the object noun phrase of the previous sentence to be treated as the subject of the following sentence, which is not a good translation. The correct phrase is "来保护环境和居民的健康。因为新冠疫情..." which translates as"To protect the environment and health of the residents. Because COVID-19..."

What might be happening is the transcriber thinks this is an instance of the 是...的construction in Chinese and places a period after the 的to end what it thinks is a complete sentence. To fix this error when it occurs, strings containing "的。" were isolated, assigned part of speech tags, and a set of rules was used to determine if there was an error or not. The part of speech tags provided information to write rules to determine if the 的。was correct, and if it was not, the period would be moved to the end of the following clause.

The last issue the markup tool solves relates to problems the transcriber commonly had with percentages and large numbers. It was quite common for the transcriptions to include things like "80%六" or "10% 17" when the speaker had said "86%" or "17%." Moreover, the transcriber didn't like percentages involving decimals very much, often writing things like "85%点七" instead of "85.7%" (点can be translated as "dot" or "point"). Issues not involving percentages such as "40000 400" when the speaker had said "40400" were also frequent. The markup tool includes a fix that notices and repairs these errors automatically, yielding better transcripts.

3 Markup Tool Results

For problems like automatic transcription, word error rate (WER) is a common performance metric. WER is calculated by the equation

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

S is the number of substitutions

D is the number of deletions

I is the number of insertions

C is the number of correct words

N is the number of words in the reference (N=S+D+C)

The lower the WER value, the more similar the transcripts are to the reference file, with 0.0 being a perfect score.

To get a reference file, a native speaker of Chinese listened to two podcasts (about 30 minutes of audio) and corrected the transcriptions by hand. These two new transcripts represent a gold standard that the original transcripts, and the transcripts passed through the markup tool, can be compared to.

For the first podcast, the original transcript had a WER of 0.0925 while the transcript passed through the markup tool had a WER of 0.0882, a relative improvement of 4.6%. For the second podcast, the original had a WER of 0.1104, while the transcript passed through the markup tool had a WER of 0.1043, an improvement of 5.5%.

However, because for Chinese word boundaries can be ambiguous, character error rate (CER) is likely a better metric. While powerful technologies already exist that tokenize Chinese texts very well, ambiguities still persist. The CER equation is the same and simply compares individual characters rather than words.

When calculating the CER, the for the first podcast, the original transcript had a CER of 0.055, the transcript passed through the markup tool had a CER of 0.054. An improvement of 1.8%. For the second podcast, the original transcript had a CER of 0.0752, the transcript passed through the markup tool had a CER of 0.0703. An improvement of 6.5%.

While these improvements are quite small, this is not necessarily a problem for two reasons. First, the intention of the program is not to make one or two specific transcripts perfect. Rather, the intention is to slightly improve any transcript and therefore improve the overall quality of the entire project. Second, the scores for the original and marked-up transcripts were both quite low, meaning the Google's Speech-to-Text API is already relatively accurate. Character and word error rate are numeric metrics that simply quantify how many difference there are between two texts, they don't say anything about the quality of the translations or summaries downstream.

Consider the following are example translations:

Original: For example, the top ten flower producing countries in terms of export revenue in 2019 are the Netherlands with as much as 46.1 US dollars, Colombia with 14.1 US dollars, Ecuador with more than 8.7 US dollars, and Kenya has a lot of reputation.

Marked up: For example, in 2019, the top ten flower producing countries in terms of export revenue were the Netherlands with US$4.6 billion, Colombia with US$1.4 billion, Ecuador with more than US$8.7, and Kenya with more than US$7 billion.

Original: Reduced social interaction during love affairs also hit flower production.

Marked up: Reduced social interaction during the pandemic also hit flower production.

And because this project aims to produce summaries in addition to translated transcriptions, it is necessary to include example summaries as well. What follows are the automatic summaries (as produced by Hugging Face's Text Summarization Transformer, the current state of the art).

Original: French people don't know that some countries have intensive industrial production of roses . The use of pesticides such as pesticides banned in Europe harms the health of workers . Other chemical elements pollute local rivers, which requires air transport abroad . The European Union has banned 41 kinds of harmful or carcinogenic Pesticides and pesticides .

Marked up: Some countries have intensive industrial production of roses, and the use of pesticides such as pesticides banned in Europe harms the health of workers . Ethiopia in Africa is the world's main exporter of roses . The European Union has banned 41 pesticides that are harmful to the human body and even cause cancer .

4 Lexical Sophistication

Something that could be useful to teachers/learners searching for language study materials is a measure of lexical sophistication. "Lexical sophistication, in turn, has to do with the use of words that are not among the most frequent in the language (e.g., ask vs. request; poor vs. destitute), and which are therefore assumed to reflect more advanced levels of vocabulary knowledge (e.g., Linnarud, 1986)"(Jarvis, 2013). This could be useful as a measure of the relative complexity of the text. There is a great deal of research in quantifying lexical

sophistication for English, with many different formulas and a good understanding of what features of a text should be considered. Features such as the frequency, range, n-gram frequency, and various psycholinguistic properties of the words of a text are all variables that should be considered (Kyle, 2014). Unfortunately, I was not able to find substantial research on the topic for Chinese, much less preexisting software that could calculate a score of lexical sophistication for us. To adequately calculate lexical sophistication for Chinese, the same formula could be used as is used on English texts. In addition to the factors listed above, factors like the complexity of individual Chinese characters should be considered (Sung et al., 2015).

5 Chinese NER

Many of the homophone issues discussed above related to named entities within the text. things like Amazon or COVID-19. A machine-learning approach to resolve the issue was attempted, setting up Chinese word embeddings using word2vec model trained on a corpus of Chinese news articles. There is research to suggest this is a good approach to named entity extraction (NER), and that information could be used to address the homophone issue (Liu et al., 2022). While the word embeddings were very accurate, the data from the trained model was not as useful as I had hoped. A machine learning approach could be very useful in improving the transcriptions, but to resolve the homophone issue the model would need to consider pronunciation, not just text.

6 Further Work

Improving speech-to-text software is an exciting and ongoing field of research. The challenge is largely acoustic, relating to phonetics and phonology. Each language presents its own problems due to the uniqueness and intricacies of each language's phonological system. Current state-of-the-art models are machine learning approaches, and improving them further is also within the purview of machine learning.

For improving the markup tool, a machine-learning approach has the potential to further polish the transcripts by repairing the homophone issues (as mentioned above). For the issue termed "的sentence breaks," to more accurately and consistently identify and resolve that error broader analysis is necessary. Constituent parsing, i.e. analyzing the syntactic structures of the text, could yield very useful information that could be used in the creation of more rules within the markup tool that can make repair this issue. For lexical sophistication, a more thorough review of the literature might make more clear what Chinese linguistic factors should be considered and how much weight should be given to each factor. From there, an equation for lexical sophistication could be developed and using a reference corpus for frequency scores a program could be made to assign scores of lexical sophistication to each podcast episode.

7 Conclusion

The internship was largely exploratory, first understanding the purpose and scope of the podcast project and then independently working to improve it. Most of the work was spent resolving creating a markup tool that resolves errors the transcription software consistently makes. Considerable time was spent setting up Chinese word embeddings using a word2vec model, but the internship was too short to find a concrete use for the software. Time was also spent researching lexical sophistication, which started because of an interest in quantifying readability of texts. Although no software was made to assign sophistication scores to texts, with a better understanding of the concept the Tech Center can more adequately approach the problem in the future.

References

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63.

Kyle, K. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49.

Liu, P., Guo, Y., Wang, F., and Li, G. (2022). Chinese named entity recognition: The state of the art. *Neurocomputing*, 473.

Sung, Y.-T., Chang, T.-H., Lin, W.-C., Hsieh, K.-S., and Chang, K. (2015). Crie: An automated analyzer for chinese texts. *Behavior research methods*, 48.