The Application of the Intelligent Learning Assistant based on Automatic Speech Recognition in the International Learning Environment

Zhongwen Yang

Xi'an Jiaotong-Liverpool University, Suzhou 215028, China

Abstract. Recently, the intelligent learning assistant is applied widely in foreign countries like America, Europe and Africa. However, the accuracy of translation and AI dialogue is not satisfied among the Chinese analogous applications due to the trait of Chinese language. To solve these problems, we propose an application which can translate and recognize user's questions. Based on Automatic Speech Recognition, we clarify the process to make it more accurate in practical use.

Keywords: Automatic Speech Recognition; Intelligent Learning Assistant; Application.

1. Introduction

With the increase of people's living standard, parents pay more attention to children's learning. More and more parents choose to send their kids to international school to study. In China, some international universities appear, for instance, Xi'an Jiaotong-liverpool University and University of Nottingham Ningbo China. The course in these university is in English and students here need to work with the foreign teachers and students. Moreover, the report from Ministry of Education of the People's Republic of China shows that there were more than 230 thousand foreign students choose to study in China in 2009 [1]. As a result, adapting in the international study became a hot topic and the language barrier is the core of it. Thereupon, the learning assistant application comes into people's view.

Some applications were developed to overcome the language barrier in international classes. For instance, an application named Otter could automatically capture meeting notes and find the information students need. Another example is Youdao dictionary, which can record what users said and translate it immediately. However, there are several problems when the applications are used in the Chinese environment. The document from Ministry of Education of the People's Republic of China shows that Chinese dialects are usually divided into ten major dialects [2]. It would be hundreds of dialects when considering the branch. Errors may happen when dealing with the dialects and the failing example is shown in Fig.1.

Figure 1. A failing example when translate the Chinese dialect

The Chinese content above is what the app record from a user, who speaks dialect. The real meaning is totally different from what the app translated. Another problem is that the translation is rigid, which means it translate the content word by word, but not the whole meaning. Moreover, the translation would be difficult if the user's grammar is not good.

To solve these problems and to make the international students have a better learning environment, we come up with an application. This app base on the speech recognition system (ASR), which can record what users say, recognize and compare it with the words of the thesaurus, whether it is Chinese, English, Japanese or other languages. The system can convert their language into the corresponding text, and translate into the desired language. In addition, this app collects common academic questions in advance, in order to provide several most suitable problems when user asks questions but cannot describe it accurately by comparing with the problems in the database. The user can input the keywords and the system would provide the related questions.

The main part of the app is to convert speech into text, which based on automatic speech recognition (ASR). Traditional word error rate is able to check the accuracy in miscues recognition, however, it cannot measure the accuracy in miscues detection [3]. As a result, to compute the latency before reading a word and to detect the miscues, the application needs to use the following methods to define and measure the accuracy of its listening [4]. We detect miscues by matching the hypothetical output of the ASR to the displayed sentence. At the highest level, which we call 'text space', we consider miscue detection as a classification problem: classify each word as read correctly or not [5]. By following this strategy, we are able to find the related questions in the database and provide the result.

Another part is communication between the user and the app. There are distinct languages in the practical use, such as English, Chinese, Japanese and even different dialects. We want the system could automatically identify the type of language after recording, and to achieve synchronous real-time conversion, truly achieve barrier-free communication. To achieve this goal, identifying the speaker is important. We set the app's dialogue architecture into four types: speech, silence, and time [6, 7]. The project aims to find out whether the user or the app take the floor. We consider to use a timer to help, for example, the app would backchannel after a 2-second silence. If the user clicks the mouse, whether on stop or others, the app would take the floor.

2. Models

The learning assistant finish the job based on the acoustic model and lexical model. The acoustic model is created by introducing a large speech database and using training algorithms to create statistical representations, which are called Hidden Markov Models. The lexical model, which powers predictive text and autocorrect for a language, is used to generate suggestions when use input the questions.

For acoustic model, the target group of the application is the students who are older than high school students, so the data from adults are more suitable than from kids. As a result, the project would introduce the TED-LIUM 3 dataset, a large collection of 452 hours of TED presentation by 2295 speakers, and divided it into two parts. The first part is used to train the initial acoustic model for ASR, and the second part is used for evaluating. The project would use the application to recognize the acoustic recordings and the accuracy would be calculated. To train the data, the project would use Markov Chain Monte Carlo methods, whose requirement is the specification of the prior distributions and constraints for all parameters [8]. This project would set accuracy Θn from 0 to 1, content divergence a from 0 to 2.5, and content difficulty bn from 0 to 1. While collecting and centralizing the trained data, this project considered the protection of user privacy. The speaker's gender and identity could be detected. The acoustic model dispersed to many devices and the fine-tune consists in the local user data [9]. When fine-tune happens, it is found that the leak of user information is possible, even without the transmission [9]. How to protect the privacy of speakers is the future topic for the project.

For lexical model, after the application convert what speaker said to the text, the project would autocorrect the text. The project wants to find different result of accuracy based on the different times user study, namely finding the learning curve, by using the method learning decomposition. According to the accuracy and simplicity, the project would choose exponential curves rather than

power curves [10]. The project creates two variables t1 and t2. T1 stands for the number of learning opportunities when the user used the application first time, T2 stands for the number of practice opportunities when user has already used the application that day [10]. The project would also estimate the parameter B which represents the relevance between the first-time user use the application and the later use, after calculating if B>1 then the learning opportunities of t1 are better; if B<1 then vice versa; if B=1 then neither is better [10]. It is found that a short sentence is hard to detect the miscue, and distracter strategy would increase the false alarm rate although it increased the miscue detection rate [4]. As a result, the project would not use this strategy. The project would use Bayesian for evaluating the relationship between the question student ask and the questions in the database, providing the best-related questions for students. For instance, when a student's question contains AI, model and sound, what are the possibilities different models contain.

3. Educational Data Mining and Experiment Estimation

Data mining has the potential capacity to collect information which is valuable to users and even the app – information could make the education more efficient, and effective so that the users' individual needs could be satisfied [11]. Through these data, we can not only optimize the performance of the software, but also achieve personalized functions for different users to improve user experience. This section analyzes and discusses two methods of data mining related to intelligent learning assistant, which are the keywords in the questions raised by users and the subsequent actions of users for different types of questions.

The app considers to collect users' information when registering the account because importing users' data is benefit for the app's interaction while some data may not be able to observe, such as gender [12], IQ [13], prior declarative knowledge [14], or pretest scores. Students' majors in education are also useful information because different majors have their unique professional vocabulary, whose frequency in the questions students asked is quite high. By collecting and analyzing the student's spoken questions to collect keywords that distinguish the different learning domains, the vocabulary in the question can be identified according to the user's professional field in the future use process, so as to improve the accuracy of speech recognition [11]. Importing keyword can also achieve the requirement based on different majors as alternative answers that are more in line with user needs can be given according to the field of expertise. The technology also faces the issue that the questions students ask may not be professional or specific as their level of language and major is not high.

The strategy to solve the issue is to reduce the range of the question by analyzing the type of questions. We can first classify questions, for example, classify questions by leading words. When user A asks a 'what... ' question, he would then go straight to the advice given by the intelligent learning assistant and ask the next 'Why... ' question. If this procedure happens frequently, we can reduce the number of alternative answers given and increase the number of possible next questions. In the other case, user B asks the question 'Why... ', but instead of choosing the answer given by the intelligent learning assistant, he repeats this question. In this circumstance, when user B asks 'Why... ', the intelligent learning assistant tends to provide more approximate answers to choose from. As students are affected by the language barrier, they don't know how to ask a proper question, or they don't know how to express themselves, so it is necessary to provide alternative answers so that users can have a wider range of choices to achieve the purpose of use.

After the operations above, we estimate the accuracy of the models and choose the degree of the model's adaptation and the efficiency of the translation as the criteria. The adaptation criterion can show the range of the potential users and the efficient criterion can show the users' using experience as the translation is the core of the international study.

To evaluate the adaptation criterion, we consider to conduct two experiments based on the gender and the noise. Considering to gender, the difference between male and female would produce an affectation to voice quality, where acoustic Voice Quality Index and Dysphonia Severity Index are

the two most widely used index to measure it [15]. However, an experiment in India shows that gender has no affectation to the acoustic indexes [15]. The result shows that DSI, AVQI, CPPs, HNR, SlopeLTAS and TiltLTAS followed the normal distribution ($p > 0.05$) in for both the genders. The parameters MPT, F0-High, I- Low, Jitter%, Shim Local and Shim dB did not follow normal distribution ($p < 0.05$) [15]. The chart is shown in Fig.2.

Table 1 Mean, Standard Deviation, Median, Interquartile range, Lower and Upper Bound across gender

Parameter	Gender	Mean (SD)	Median (IQR)	95% Confidence interval for mean	
				Lower bound	Upper bound
AVQI	Male	1.79 (0.70)	1.85 (1.10)	1.61	1.96
	Female	1.87 (0.59)	1.86 (0.84)	1.87	2.01
CPPs	Male	16.68 (1.47)	16.65 (1.53)	16.31	17.05
	Female	14.89 (1.15)	14.78 (1.67)	14.62	15.15
HNR	Male	20.98 (2.1)	20.99 (2.65)	20.44	21.52
	Female	23.26 (1.63)	23.26 (2.47)	22.88	23.64
ShimLocal	Male	3.53 (0.97)	3.47 (1.29)	3.30	3.79
	Female	2.69 (0.65)	2.57 (0.99)	2.54	2.84
ShimdB	Male	0.37 (0.072)	0.37 (0.08)	0.35	0.39
	Female	0.30 (0.053)	0.30 (0.08)	0.29	0.32
SlopeLTAS	Male	−20.42 (3.44)	−20.11 (4.08)	−21.28	−19.57
	Female	−20.30 (4.91)	−20.82 (6.62)	−21.44	−19.16
TiltLTAS	Male	−11.18 (0.89)	−11.24 (1.31)	−11.40	−10.95
	Female	−10.82 (0.94)	−10.79 (1.29)	−11.04	−10.61

Figure 2. The result of the experiment [15]

Another potential factor we considered is the noise. The lecture room is the main place we want to deal with as this place contains a lot of students. We want to examine that is the application able to recognize users' speech input or the lecturer's speech input. We would record the noise level with different quantity of students in the lecture room and use the distinct decibel as the matched group.

4. Summary

International study is a popular learning pattern and there are lots of students who get into trouble with the language barrier. Many applications are used to help for the oversee students, however, the Chinese current applications, which remain problems with the translation and artificial intelligence dialogue, are not powerful enough. As a result, we propose the learning assistant, which is suitable for the Chinese learning environment and has a wider range of application, based on ASR. Meanwhile, we set the related experiments with the criterion of the models' adaptation and the translation efficiency to certify the effectiveness of the models. In the future, we may need the further exploration at the application's accuracy and expandability.

References

[1] "In 2009, the number of Chinese students studying in China exceeded 230,000," 22 March 2010. [Online]. Available: http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/moe_1485/201005/t20100518_88315.html.

[2] "An overview of the Chinese language," Ministry of Education of the People's Republic of China, 27 August 2021. [Online]. Available: http:// www. moe. gov.cn/ jyb_sjzl/ wenzi/ 202108/ t2021 0827 _554 992. html.

[3] J. Mostow, "Is ASR accurate enough for automated reading tutors and how can we tell?," Pittsburgh, PA, 2006, pp. 837-840.

[4] J. Mostow, "Why and How Our Automated Reading Tutor Listens," Carnegie Mellon University, Pittsburgh.

[5] A. G. H. L. L. C. a. S. R. J. Mostow, "Towards a reading coach that listens: automated detection of oral reading errors," Washington, DC, 1993, pp. 392-397.

[6] J. M. G. Aist, "A time to be silent and a time to speak: Time-sensitive communicative actions in a reading tutor that listens," in AAAI Fall Symposium on Communicative Actions in Humans and Machines, Boston, MA, 1997.

[7] G. Aist, "Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption," in Proceedings of the International Conference on Speech and Language Processing (ICSLP98), Sydney, Australia, 1998.

[8] J. M. Yanbo Xu, "A Unified 5-Dimendional Framework for Student Models," Pittsburgh, PA.

[9] S. B. J.-F. T. M. T. N. E. Y. Mdhaffar, "Retrieving Speaker Information from Personalized Acoustic Models for Speech Recognition," in Acoutics, Speech and Signal Processing (ICASSP), Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2022-2022 IEEE International Conferenc, 2022.

[10] J. M. Joseph E. Beck, "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practive for different types of students".

[11] J. B. Jack Mostow, "Some useful tactics to modify, map and mine data from intelligent tutors," Pittsburgh, PA, 2005.

[12] B. W. B. a. S. Arroyo, "Macroadapting AnimalWatch to gender and cognitive differences with respect to hint interactivity and symbolism," in 5th International Conference on Intelligent Tutoring Systems (ITS2000), Montreal, Canada, 2000.

[13] G.-G. Y. B. Shute, "An experiential system for learning probility: Stat Lady description and evaluation.," in Instructional Science 24(1), 1996, pp. 25-46.

[14] M. S. Corbett, "Modeling student knowledge: Cognitive tutors in high school and college," in User modeling and user-adapted interaction 10: 81-108, 2000.

[15] S. P. M. Shabnam, "Effect of Gender on Acoustic Voice Quality Index 02.03 and Dysphonia Severity Index in Indian Normophonic Adults," 2021. [Online]. Available: https://doi-org.ez.xjtlu. edu.cn/ 10. 1007/ s12070-021-02712-8.

[16] J. B. Jack Mostow, "Some useful tactics to modify, map and mine data from intelligent tutors," Pittsburgh, PA, 2005.