Design of Annotation system of Russian Speech Corpus Based on Crowdsourcing

YanzhouMa[1, a]

[1] University of Foreign Languages, Luoyang, Henan 471003, China

[a]mypollywell@163.com

Keywords: Russian Speech Corpus,Annotation system,Crowdsourcing.

Abstract. Transliteration annotation in Russian speech recognition corpus has been always a problem which is difficult to resolve due to its time consuming, low efficiency and difficulty in personnel and quality control. This paper presents a technology based on crowdsourcing, designs the crowdsourcing platform within LAN and calls Russian majors to finish the large-scale corpus transliteration annotation task in a short time. After analysis and comparison, the results are comparable to those of the traditional methods. It provides a reference for improving the efficiency of speech corpus transliteration annotation and puts forward suggestions on improvements of some problems.

1. Introduction

With the further development of speech technology research, not only the text corpus is needed to build the language model, but also the large-scale speech corpus is required to construct the acoustic model. Compared with the text corpus, the speech corpus can record the spectrums of speeches and relevant parameters of acoustics in details, and annotate the syntax, rhythm and other information related to language, so it is widely used in speech recognition and other relevant research fields. A key problem Russian speech recognition [1] faces is the training of the language model and speech model which are correlated with the corpus. The selection and application of corpus follows the principles of representativeness and coverage, but the annotation work of Russian speech corpus[2] resources has been always a difficulty consuming both time and energy. Adopting the crowdsourcing technology will make it become simple and efficient with the reasonable task distribution and appropriate incentive mechanism. The speech library annotation can be good compatible with other disciplines. Many aspects of crowdsourcing are suitable for the speech library annotation research. With the crowdsourcing technology, researchers can explore many new research approaches. A large number of studies have shown that using the crowdsourcing technology can finish the large-scale speech library collection and annotation of relevant data in a short time

On the basis of the crowdsourcing technology, this paper studies the construction process of Russian speech recognition corpus [3]. It focuses on the role of the crowdsourcing technology in the annotation process of speech corpus and design and realization of the online transliteration annotation platform based on crowdsourcing and discusses how to effectively develop the construction period and the quality control issues. Finally, through analysis and comparison, it proposes the solution of the speech library construction based on crowdsourcing.

2. Basic Concept and Method of Crowdsourcing

2.1 Basic Concept of Crowdsourcing.

The concept of crowdsourcing was proposed by Jeff Howe, a reporter of Wired in the US in June 2006. He defined crowdsourcing [4] as: “a practice that a company or organization outsources the tasks which were done by the workers in the past to non-specific (and usually large) public network. The tasks of crowdsourcing are usually borne by individuals, but if more people are required to complete a task, it may appear in the form of the open source individual production.” The basic characteristics of crowdsourcing are: (1) it calls the Internet public in an open way; (2) the crowdsourcing tasks are usually the problems which are difficult to resolve by computer alone; (3) the

public complete the tasks collaboratively or independently; (4) it is a distributed problem solving mechanism.

Since its rise in 2009, the crowdsourcing technology [5] has been praised by a large number of scientific research personnel. It is mainly applied in the following fields: database, natural treatment, machine learning and artificial intelligence, in which speech recognition is just included. The collection and annotation of Russian speech recognition corpus [6] is a complicated and time-consuming task. The most ideal way is to adopt the open way to call the public to complete it. In addition, the collection and annotation task is difficult to be solved by a single person and a single computer. Adopting the crowdsourcing technology to complete the task collaboratively or independently is completely reliable and it is an effective solution mechanism to the work issue. Russian speech recognition requires the large-scale speech corpus which needs over 500 h to train an ideal acoustic model. Building a 500-h Russian speech library with annotations needs about 5000 h to annotate the speeches and the annotation personnel should grasp the Russian syntax and grammar knowledge. If each person works 8 h per day, it needs 1 person to work for 625 days or 10 persons to work for 62.5 days to complete it. If adopting the crowdsourcing technology to complete the 500 h of annotation work, it can be finished in a short time. If each person works 2 h, it needs 10 persons to work for 250 days or 50 persons in 50 days to complete the task. Combined with the effective quality control strategy, it can achieve the purpose.

2.2 Basic Flow of Crowdsourcing.

Main participants of crowdsourcing [7] include task requesting people and task executors. They are linked together through the task. The flow is shown in Fig.1:

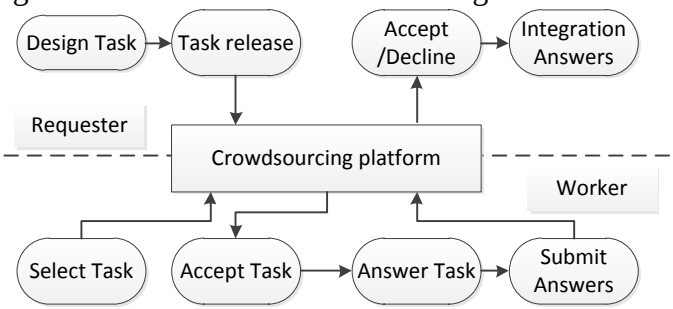

Fig.1 Crowdsourcing Work Flow

As shown in Fig.1, when the task requesting people plans to use crowdsourcing to finish the task, he needs to use crowdsourcing [8] in the following steps: 1. design the task, 2. release the task with the crowdsourcing platform, 3. wait for the solution, 4. refuse or receive the answers from the executors, 5. consolidate answers, and 6. complete the task. The executors need to do in the following steps: 1. look for the interesting task, 2. receive the task, 3. finish the task, and 4. submit the task. From the time dimension, the working process of crowdsourcing can be divided into three stages: task preparation, execution and answer consolidation, in which task preparation includes task design and release by the task requesting people and task selection by the executors, task execution includes task receiving, solving and submitting by the executors and answer consolidation includes: accepting or refusing the answers and consolidating the answers by the task requesting people.

2.3Research on the Key Problems in Crowdsourcing.

(1) Task Preparation

It plans to collect and annotate 500 h of Russian speeches. The target speeches (fields, regions, speakers, speakers' sex and age) are divided into different price systems according to the difficulty of the task. Many factors contribute to the result quality of a task, such as the difficulty of the task, the responsibility of the executors, the correlation between the task and the executors. In addition, since it is difficult to ensure the result quality of the task by relying one executor's answer, the task requesting people always assign the task to many people and deduce the final result with different strategies in the answer consolidation stage. Through the description of the task interface, people who finish the task can obtain the specific information about it and completely understand the relevant requests, so the design of the task interface is very important.

(2) Task Selection

The research focus of task selection is how to help the executors select the tasks related to them. From the perspective of the task discoverer, there are two main task selection methods: the pull-based method and the push-based method. In the first method, the executors take the initiative to search related tasks, namely, task searching; in the second method, the crowdsourcing platform assigns the task, namely, task recommendation.

The online transliteration annotation platform [9] designed in the paper supports the task search ability. The executors browse the tasks to find the information they are interested in. The task recommendation method does not need them to input the query information actively. Instead, the crowdsourcing platform takes the initiative to recommend the tasks according to different executors' interests. The historical records reserved on the crowdsourcing platform are the best reflection of the performance of the executors.

(3) Task Execution

The main challenge in the task execution stage is how to effectively combine the factors of persons involved and the task optimization purpose of the requesting person to assign the task online. In the task execution process, assigning the task to appropriate persons through the effective online task assignment strategy can improve the quality of the task results. Result evaluation and replacement strategy are adopted to realize the dynamic task allocation. The task requesting people just assign a part of tasks at the beginning, then replace the people with low quality of the results by the replacement method according to the feedback answer evaluation and the ability of the works, and finally reassign the tasks which have been finished before, so as to improve the quality of the results.

(4) Answer Consolidation

Due to the different ages and educational backgrounds, the executors involved are inevitably affected by the subjective consciousness and knowledge background during answering the questions. Therefore, the quality of the answers changes greatly when finishing different types of tasks. To ensure the result quality of the task, different control strategies are put forward. Improving the quality of the result according to the rate of accuracy in problem solving is to score each executor according to the rate of accuracy. Those who gain higher scores are given more weights. Finally, the answers provided by the executors are weighted to determine the final result according to the weighted point.

3. Design and Development of the Crowdsourcing Platform.

The business process is mainly divided into two parts. In the first part, the administrator logs in the system and enters the background management module to realize the control of the resources. In the second part, the common users log in the system and participate in different corpus annotation tasks. All users and administrators need to log into the system [10]. If it is the first time for common users to log into the system, they first select the tasks they are interested in and then participate in the corresponding ability test. If they pass the test, they can enter the task interface to complete the tasks. If they fail to pass the test, they must continue to do it until the rate of accuracy reaches the standard. After logging into the system, the administrator realizes the management of the system resources and the users through viewing the relevant information in the system [11]. The business flow chart of the system is shown in Fig.2.

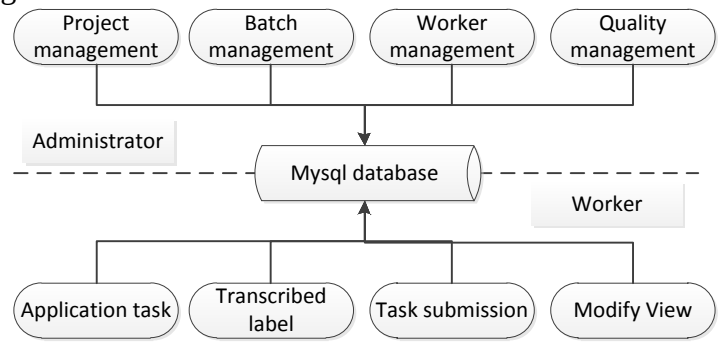

Fig.2 Framework Structure of the Transliteration Annotation System

Administrators and task applicants access the background database through different interfaces. Administrators make the project management, batch management, personnel management and quality management. Project management includes creating contents, adding project properties and instructions, setting the default values of project properties, importing labels and confirming the release. Batch management includes creating contents, importing the original Russian speech files, importing texts and confirming the release. Task management includes checking the task list under a link, querying specific tasks, querying the task of a speech number and inspecting the quality through screening. Personnel management includes editing the basic information of applicants, checking all tasks of the applicants and confirming using the applicants or not. Quality management includes specifying the quality supervisor, distributing the authority of inspection personnel and checking the work details of quality supervisors.

Task applicants enter the task interface according to the assigned accounts and passwords, query the tasks in the project list and apply to join a project. The platform agreement will pop up, regardless of random application or direction application. If the applicants select to work, they will see the information related to the task, instruction of the task operation area and annotation area of the sound attributes and can do transliteration annotation. In the task operation area, they carry out the annotation of characteristics and attributes of speech files and characteristics of text contents, transliteration annotation of the contents corresponding to texts and speech files. After completion, it comes to the next speech file and be saved automatically until the completion of the applied task.

The transliteration annotation platform [12] is based on PHP development and deployed in the Linux environment. It requires Centos6.5 version or above. MySQL requires 5.4 version or above.

4. Quality Control of Crowdsourcing

Quality control is an extremely important step in the design process of the crowdsourcing system [13]. Traditional quality control methods which are frequently used, such as monitoring mechanism, social norms and contract signing, are not very effective for the tasks on the Internet. Therefore, on the premise that no any external mechanism can improve the result quality of tasks, the best way is to turn to the task setting itself.

4.1Keeping simplicity in task setting.

Crowdsourcing relies on mass contribution to complete the task, but it is easy for the executors to make mistakes. Then often provide some inaccurate results for two reasons. First of all, in order to get more benefits, the executors provide some random results for all problems, which will greatly affect the quality of the result. Secondly, for some complex tasks, the executors lack the knowledge to handle them, which may lead to the cognition deviation and then incorrect results. To solve the above problems, it is necessary to set the tasks to be simple and clear enough, so that the executors will not feel difficult or lose patience due to the complexity. A task can be divided into many micro-tasks which are assigned to different people. Each micro-task can obtain more than one result. Finally, the correct result can be selected through developing the reasonable strategy.

4.2Adding the test process for each task.

Task releasers may worry that the users provide some irresponsible random results. A common method is to set some questions related to the theme of the task to make a preliminary evaluation and elimination of the participants. The premise is to ensure the testing questions are accurate. Not every user entering the system are suitable for participating in the task practice, so it is necessary to eliminate the useless users and make full of the high-quality users. Therefore, to ensure the quality of the obtained data, the qualification examination is introduced. The so-called qualification examination means testing the ability of the users when they enter the system. After passing the examination, they can do the task. If the users pass the test link of the task, the system will automatically generate the task interface link. The users can select entering the task practice or continuing the test.

4.3Auditing Strategy of the Annotation Results.

In the process of quality control, administrator auditing is also very important. He controls the resources through managing the data in the system. The professionals in the Russian research field as

the administrator can audit the annotation results directly. The administrator checks the annotation records of the system tasks, evaluate the annotation results of the users, keep the qualified annotation results and removes the junk annotation results. He can also assess the level of users through checking the historical records of the users and give ordinary and experienced users different evaluation weights to let experienced users have more system permissions and the weights of decision-making. After the weighted calculation of the scores of each record, he can ultimately determine the quality of each record.

5. Summary

This paper proposes the technology solution for the construction of Russian speech corpus based on crowdsourcing and designs the crowdsourcing-based transliteration annotation platform. Under the support of the platform, it takes 55 days to finish 500 h of the transliteration annotation work of speech corpus. Compared with the traditional method, it has great improvements in time consumption, efficiency and quality in the corpus annotation. Through comparative analysis, the annotated corpus has the same quality as that obtained by traditional methods. However, there are also some problems, such as too single incentive mechanism of the system and more Russian students as the users. It will be further improved through further expanding and the numbers of concurrent users and the resource annotation tasks and checking the efficiency of the system.

References

[1] Ronzhin AL,Yusupov RM,Li IV, et al. Survey of Russian Speech Recognition Systems[C]//Specom, [S.l.]: Citeseer, 2006: 54-60.

[2] Arlazarov VL,Bogdanov DS,Krivnova OF, et al. Creation of Russian Speech Databases: Design, Processing, Development Tools[C]//9th Conference Speech and Computer, [S.l.]: [s.n.], 2004: 18-23.

[3] Karpov A,Markov K,Kipyatkova I, et al. Large Vocabulary Russian Speech Recognition Using Syntactico-statistical Language Modeling[J]. Speech Communication, 2014, 56(1): 213-228.

[4] Howe J. The Rise of Crowdsourcing[J]. Wired Magazine, 2006, 14(6): 1-4.

[5] Doan A,Ramakrishnan R,Halevy AY. Crowdsourcing Systems on the World-wide Web[J]. Communications of the Acm, 2011, 54(4): 86-96.

[6] Skrelin PA,Volskaya NB,Kocharov D, et al. A Fully Annotated Corpus of Russian Speech.[C]//Lrec, [S.l.]: [s.n.], 2010: 18-23.

[7] Parent G,Eskenazi M. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges.[C]//Interspeech, [S.l.]: Citeseer, 2011: 3037-3040.

[8] Freitas J,Calado A,Braga D, et al. Crowdsourcing Platform for Large-scale Speech Data Collection[J]. Proc. Fala, 2010, (2): 19-25.

[9] Su AY,Yang SJ,Hwang W, et al. A Web 2.0-based Collaborative Annotation System for Enhancing Knowledge Sharing in Collaborative Learning Environments[J]. Computers & Education, 2010, 55(2): 752-766.

[10] Dowell RD,Jokerst RM,Day A, et al. The Distributed Annotation System[J]. Bmc Bioinformatics, 2001, 2(1): 7.

[11] Sakata K,Nagamura Y,Numa H, et al. Ricegaas: an Automated Annotation System and Database for Rice Genome Sequence[J]. Nucleic Acids Research, 2002, 30(1): 98-102.

[12] Volkmer T,Smith JR,Natsev AP. A Web-based System for Collaborative Annotation of Large Image and Video Collections: an Evaluation and User Study[C]//Proceedings of the 13th Annual Acm International Conference on Multimedia, [S.l.]: Acm, 2005: 892-901.

[13] Adda G,Mariani JJ,Besacier L, et al. Economic and Ethical Background of Crowdsourcing for Speech[J]. Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, 2013,(6): 303-334.