Research and Implementation of the Basic Corpus Management System

Maobo An [1, a], Yuan Huang [1, b], Xiaochen Sun [1, c], Shengxiang Gao [1, d], Xin Jin [1, e] and He Gao [2, f]

[1]National Computer Network and Information Security Management Center, Beijing 100029, China

[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

[a]maoboan@126.com, [b]huangyu698@126.com, [c]sunxiaochen689@163.com, [d]gaosx689@163.com, [e]jinx259@163.com, [f]gao_he@bupt.edu.cn

Keywords: Corpus; Corpus processing; Corpus annotation.

Abstract. Basic corpus can be used as an optimization AL recognition engine of system speech and performance testing from recognition technology, the main implementation of the basic corpus is the fine-grained annotation of the language, the speaker, the channel and the content of speech. In this paper, we do extensive research on basic corpus and implement a management system, which can be used to inquire, count and operate the data of corpus, and supporting the data operations such as diplacusis, annotation, modification and exportation on the speech data. Our system also realized the systematic, standardized and structured management of the basic corpus data.

1. Introduction

Speech corpus is a collection of speech data and its annotation for the speech technology research and development. Since the 1980s, study on the development and application for corpus under the strong support of computer technology and made great progress. It had established multi-lingual speech corpus one after another, most of them are based on English, so large-scale Chinese speech corpus is very important to the research and development for Chinese speech processing technology [1].

Since the 1990s, there are more than dozens of universities and research institutes have extended the construction and study for Chinese speech corpus in our country [2, 3]. The company of IFLYTEK from USTC issued Chinese speech corpus which scale was over the size of 2.7GB voice data, including male and female, with a manual and automatic labeling standard method combination, had been applied to IFLYTEK speech synthesis and recognition systems. Microsoft Research Asia established the Chinese corpus including about One hundred and eighty thousand syllables and mainly used in the Chinese rhythm analysis and speech synthesis. Chinese Academy of Sciences, Tsinghua University and Peking University had already established a mandarin speech corpus which is mainly used for speech analysis and synthesis of research.

2. Summary of Corpus

At present, the foundation of large-scale corpus and the research based on corpus is one of the linguistics research trends at home and abroad [4]. Speech corpus contains text and voice libraries and based on voice fact, through natural voice acquisition, voice annotation, retrieval, statistical and other functions. Establish a speech corpus is the foundation of the study of the voice. From an engineering perspective, the speech corpus is one of the important part of voice engineering and the foundation of the voice system.

There are many types of corpus, the main basis for determining the type of corpus is the purpose of research and use, this point is usually able to reflect on the principles and methods of data collection [5]. Someone once divided the corpus into four types: 1) Heterogeneous: There is no specific data collection principle, it's widely collected and stored a variety of materials; 2) Homogeneous: Only collect the same type of content data; 3) Systematic: According to the pre-determined principle and percentage collecting the corpus to make the corpus balanced and systematic which can represent a range of linguistic facts; 4) Specialized: Only collect the corpus for a particular purpose. In addition, according to the corpus of language, the corpus can be divided into monolingual, bilingual and multilingual [7]. According to the corpus collection units, it can be divided into discourse, sentence and phrase. Bilingual and multilingual corpus in accordance with organizational forms can also be divided into parallel corpus and comparable corpus, the former constitute translation relations and more used in machine translation, bilingual dictionary compilation and other applications, the latter will express the same content but different language text which will be collected together and more used in the contrastive study of the language.

3. Corpus Management Scheme

The function of a computer corpus is mainly related with three factors. First is the scale of corpus, the second is distribution of corpus and the third is the degree of processing data. The scale's size is related to the reliable of statistical data, the distribution of corpus refers to the application scope of statistical result, the depth of processing data determines what kind of linguistic information the corpus can provide to users [8].

The corpus management software implements data acquisition, processing, management, etc. It provides friendly interface to the user and supplies speech data for a variety of voice analysis engine optimization.

The system mainly includes the functions in the Fig. 1.

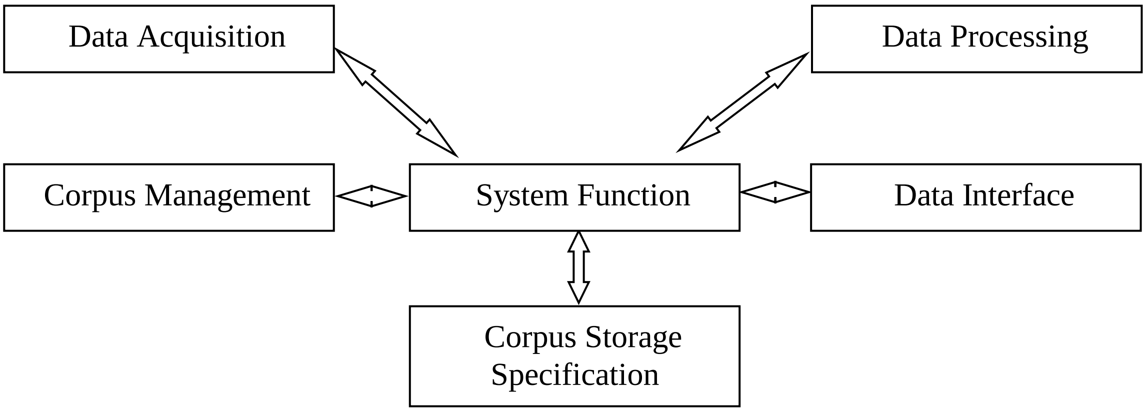

Fig. 1 System function

3.1 Corpus Acquisition.

The system will automatically obtain corpus information from specific data source, find the new adding voice files the day before and copy these files to the classification storage of speech corpus. The storage is divided into two parts, one is database storage, in the form of database records stored the beginning time of the voice ringing, the ending time of the voice ringing, the starting time of the calls, the terminal time of the calls, the calling number and the called number. Another is the file storage, which stored in the form of speech files on the server and stored classification according to the time and source.

3.2 Corpus Processing.

Corpus processing refers to the data acquisition will extract including text information, recording and the processing of preservation through the artificial complex listening or transcribing process automatically by computer program. Speech information from the corpus is divided into audio files and data text files and there is a one-to-one correspondence between them. Audio files is the base of the corpus, the simple formats of audio files are common, such as PCM,MP3 or WAV, the length of audio file is always more than 15 seconds. Data text file is the content file of audio corresponding to audio file, including the length of audio file, the starting time of each speech contents, the content of speech text, speaking people, the speaker gender, the classification of language and so on.

3.3 Corpus Management.

After scientific selection and tagging the appropriate scale corpus, there should be a full-featured management system, the system should include data maintenance (data entry, proofreading, storage, modification, removing and data description information management), corpus processing (word segmentation, tagging, text division, merging, corpus alignment and tag processing), the function of customer services (query, retrieval, statistics, printing and so on). The data maintenance part mainly relates to Chinese characters processing, text processing, file management and other computer programming technology. The main content of corpus automatic processing part is automatic word segmentation and annotation of various linguistic attributes [6].

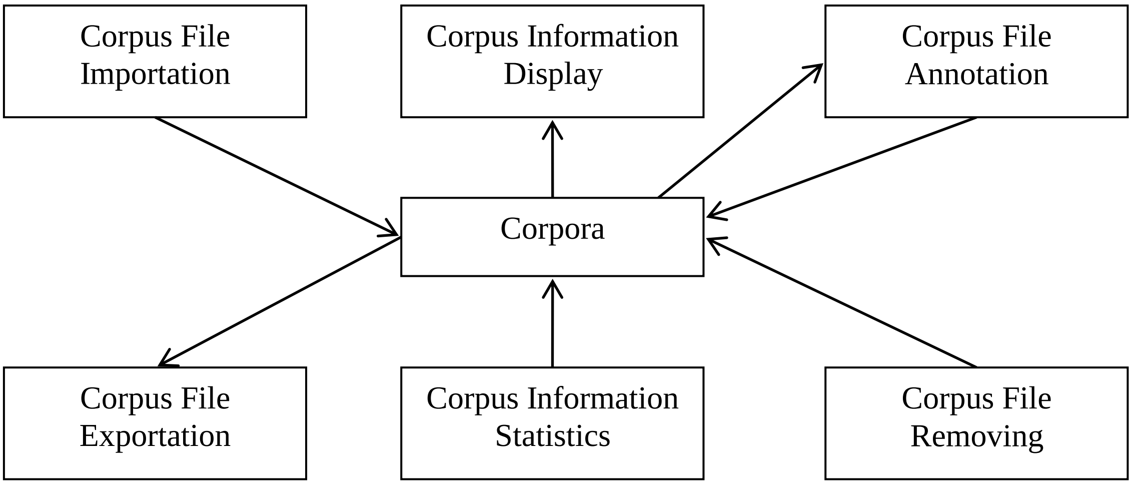

Fig.2 Corpus management system

Corpus Information Display. The speech information in the corpus displayed in the form of a list, it contains speech keyword, events, time, subject, annotate or not, the people of annotation, the time of annotation, removing and view.

Corpus File Annotation. The speech information in the corpus displayed in the form of a list, it contains speech keyword, events, time, subject, annotate or not, the people of annotation, the time of annotation, removing and view.

Speech tagging interface is divided into two layers: The upper is the interface for speech playing (including an adding button for annotating layer). The lower is the interface for annotating. Annotation layer interface display the serial number, the name of annotation layer, the starting time and ending time for annotation by vertical list. When you select the name of the annotation layer, the layer will display the name of annotation, the time of annotation, the people of annotation and operating button (including playing, removing and saving) in the tab.

Clicking the adding button of annotation layer, it will add a new tab directly in the interface and the user can input the name of the layer, the starting time of annotation, the ending time of the annotation and the content of the annotation. You can drag the playing area and automatically fill the starting time of annotation and the ending time of annotation, you can also manual input identified selecting annotation layer. When the user is annotating, he or she can select only the playing area which is selected annotation. When the user clicks the button of saving, the page will judge the range of annotating time whether conflict to the existing range of annotating time, if there is a conflict, the user will be prompted to modify the time range. The new annotation tab is in the top of the other tabs by default, after saving the annotation successful by the user, the page will refresh automatically, from top to bottom arranging the order of layer tab according to the time.

You can directly modify the name of the annotation, the content of the annotation and the selected area after you select the layer. The starting time of selected area must be greater than or equal to zero or no less than the last marked ending time, the ending time of selected area must be less than or equal to the ending time of the playing speech or no more than the next starting time of annotation.

The user can select the tab of annotation and click the deleting button, and after that the selected annotation will be deleted, the corresponding annotation will be deleted from the database too.

If the speech file has been marked with more annotation layers, when you are playing the speech, the lower layer of interface will automatically show the annotation layer and its content according to the processing of playing time. The user can view part of or all of the tagging content according to playing processing.

Corpus File Importation. Corpus file importation refers to users import the corpus file to systematic based corpus from external, including speech file importation and marked file importation.

Put the speech and annotation files in the same folder, when you are importing the files, the user must input the local directory he will import. When the system puts the files in the based corpus if you find there are many speech files with WAV and MP3 format, you should create a child directory in that folder and convert WAV and MP3 format of speech files to PCM format files and then store those files in that folder, the original WAV and MP3 format of speech files should store in the folder's child directory.

In order to prevent the conflict with the name of imported files, the speech and annotation files which are imported in the corpus will add underscore and timestamp for renaming based on the original files. After renaming, the speech files are corresponding to the annotation files.

The user imports the external files through the management system, only the file name, storage time, whether marked and annotation content in the list fields.

Corpus File Removing. Through the corpus management software you can delete corpus files and related information including corpus information in the database, corpus files and annotation files.

Corpus File Exportation. Corpus file exportation refers to the user export the corpus files to local computer from systematic based corpus, that include speech list files(Excel format), speech files and annotation files exportation. Users can query the list of exporting speech files according to the conditions from based speech management interface and click the export button in the interface. The content of exporting is all of the speech data and files which satisfied the user's query condition.

Corpus Information Statistics. The management system will count the corpus data automatically and store the results in the database. Statistical content includes language statistics, the speaker statistics and keyword statistics, statistics according to the day, the statistical results are also stored by the day.

Statistics for the number of corpus files: A certain period of time according to the total number of annotation files by per day, month and year.

Statistics for the number of each user's annotation: A certain period of time according to the number of annotation files by the user statistics (only show the top ten).

Statistics for the frequency of keyword: A certain period of time according to the number of annotation files by keyword statistics (only show the top ten).

Statistical results can be presented in a list, stitches diagram and chart.

3.4 Corpus Storage Specification.

Corpus files stored according to certain specifications, it is convenient to get corpus data directly, corpus storage must follow the principle that is simple structure and easy to find.

According to the source of corpus for storage. Storage classified by the producing source of the corpus, there are two main kinds of corpus sources, one is producing by the system and another is producing through other way to get.

According to the producing time of corpus for storage. Based on the source of corpus and according to the producing time of corpus for storage, it is very convenient to search.

The naming conventions. Corpus includes speech files and annotation files and the files are correspondence between each other, they are stored with the same name but different extension.

3.5 Data Interface.

The getting interface of corpus. Corpus management system automatically get the information of corpus, they are stored in the database and the file system respectively.

Corpus importing interface. The management system provides the function of corpus importing, you can import the information of corpus through the corpus management system which includes speech files and annotation files, the system will automatically parse recognition, store preservation and associate actions. It also provides the function of querying and marking with the importing corpus.

Corpus exporting interface. All the information from the corpus can be exported by the system and the system will compress the exporting files.

Speech annotation and data analysis interface. Corpus information can be provided the data analysis engine to train for the model.

4. Conclusion

The establishment and management of speech corpus is a complex problem, because the speech situation is not the same, so the specific speech corpus will encounter various difficulties in the process of the establishment and management, In this paper, we will provide a feasible scheme for the establishment and management of speech corpus, i hope that will provide reference for the research of speech corpus.

References

[1] Sen Zhang, Lei Liu, Luhong Diao. Problems on Large-Scale Speech Corpus and the Applications in TTS. Chinese Journal of Computers. Vol. 33 (2010) No. 4, p. 687-695.

[2] Lianhong Cai, Dandan Cui, Rui Cai. TH-Coss, a mandarin for speech corpus TTS. Journal of Chinese Information Processing. Vol. 21 (2007) No. 2, p. 94-99.

[3] Shengliang Tang, Shili Zhang, Zhiping Zhang, et al. Speech-synthesis system based on news broadcasting corpus. Proceedings of the 8th National Conference on Man-Machine Speech Communication, Beijing, 2005, p. 326-329.

[4] Faxin Zou: Design and Implementation of Speech Corpus (Master, Guangxi Normal University, China 2012). p. 31-38.

[5] Huizhong Yang. An Introduction to Corpus Linguistics. Shanghai Foreign Language Education Press, 2002, p. 20-35.

[6] Yingquan Shen, Yongjin Liu, Jun Cai, et al. Method and implementation of transcribing speech corpora based on human-computation. CAAI Transactions on Intelligent Systems. Vol. 4 (2009) No. 3, p. 270-277.

[7] Tingting He: Study on Corpus (Doctor, Central China Normal University, China 2003). p. 61-66.

[8] Tongxuan Zhang. Design of Folk Song Corpus Based on Web Retrieval. Modern Electronics Technique. Vol. 333 (2010) No. 22, p. 38-41.