Design and Development of Chinese Teaching Software based on Chinese Audio-visual Multimedia Corpus

Zhang Yanjun[#1]
[#]Huawen College of Jinan University, Guangzhou 510610, China

Abstract

The subtitle corpus is constructed with the multimedia audio-visual resources such as Chinese film, Chinese video program, Chinese song and so on, and the corpus is associated with Chinese character grade outline, vocabulary grade outline, culture grade outline, grammar grade outline, etc. Manually annotate, get the level knowledge point of Chinese teaching, to develop a software that can not only watch movies, but also study and test Chinese at the same time. The multimedia subtitle corpus can be multi-dimensional and vividly reflects the reality of Chinese language life. It can put the students in a broad and real multimedia situation, so that the Chinese film is not only the entertainment object of the students, but also a vivid Chinese teaching material, which can make the entertainment and study organic set, and realize the entertainment of Chinese. It is an interesting teaching and learning.

Keywords — *multimedia, corpus, outline*

I. INTRODUCTION

With regard to the construction of a Chinese education service platform, the simple construction of a platform or the development of some Chinese education resources search software can no longer solve the fundamental problem of the utilization of Chinese education resources, the Chinese education service platform has been designed again in a comprehensive way. If there is no resources, users can not access the platform, search engines can not understand and deal with the semantic and knowledge needs of users, the retrieval results are only literally in line with the users' requirements, and retrieval return content is often not available. Therefore, as the leader of the Chinese language education service platform, the development of the Chinese language Institute mainly lies in two aspects: first, the establishment of a monitoring directory of Chinese language education resources (to avoid blind search by users); and the second is the construction of Chinese language education resources with special characteristics[1].

In recent years, with the development of Chinese economy, the popularity of Chinese language is increasing, and more students learn Chinese language and culture. Chinese movies, Chinese video programs, Chinese songs and other audiovisual resources account for a large proportion of the students' study, life and entertainment. They are also consciously and unconsciously learning Chinese through these audio-visual media. In general, Chinese colleges offer courses such as appreciation of Chinese movies, singing and learning Chinese as a combination of study and entertainment. The subtitles of these Chinese media materials are a corpus that covers a wide range of contents. Compared with the "student written corpus" and "student spoken language corpus", the multimedia subtitle corpus breaks through the limitations of the monotonous format of the previous corpus. Can more dimensions, vividly and realistically reflect the state of Chinese language life. The paper is based on subtitle corpus and is associated with Chinese character rank outline, vocabulary grade outline, culture grade outline and grammar grade outline, which is bound to be a knowledgeable and interesting subject. As the leader of the Chinese language education service platform, the Chinese language Institute is bound to develop some characteristic applied products for study, so the subject is meaningful in both theory and application[2].

It is a perfect time to learn and remember when people are enjoying themselves and relaxing. A professor of Chinese at the Chinese institute is teaching "Liang Shanbo and Zhu Yingtai". In the teaching design, the story reading and PPT pictures were combined to develop the story plot, to help the students understand the main idea of the text. His teaching side got a unanimous praise from the teachers and students. In fact, it was a way of using multimedia to promote students' learning through a full range of audiovisual. The mining learning of words, culture and grammar based on subtitle corpus is also based on the same means. The mining learning based on the subtitle corpus is to put students in a generalized real multimedia situation, and the effect will be better. The multi-directional, multi-angle and vivid multimedia subtitle corpus breaks through the limitation of the monotonous format of the former corpus, and can reflect the language reality of the Chinese language more vividly and realistically. Data mining makes the subtitles associated with the outline of Chinese characters, the outline of vocabulary, the outline of cultural grade, and the outline of grammar, which makes the multimedia subtitles a vivid textbook for watching movies. Watching shows is more inclined from the entertainment side to the

learning end[3].

This topic will develop multimedia video and audio learning software which supports multilingual versions (mainly Chinese to English, Chinese to Indonesian, Chinese to Thai, etc.) and has the function of corpus mining. Realize the association mining according to "Chinese character rank outline", "vocabulary grade outline", "culture grade outline", "grammar grade outline", automatic tagging (if the machine is difficult to achieve, you can manually annotate the subtitle corpus of the film, for example, grammar, etc. and the cultural knowledge points) form subtitles of words, culture and grammar dictionaries, so that the corpus has been linked to the study of the tagging, when watching the film also know that those are knowledge points.

II. RESEARCH STATUS

The current status of foreign research is as following：

Foreign studies in this area generally appear in English, Japanese and Korean subtitles. Subtitles in English are called "subtitle", "title", "caption" and so on, by reading the English literature published in recent years related to subtitles, for example, some mainly study the language translation of subtitles and the language and culture of subtitles, and some study the language characteristics of subtitles based on database, and uses film subtitles as a corpus. There is no applied research in this field, which is related to Chinese character grade outline, vocabulary grade outline, culture grade outline, grammar grade outline and so on[4-5].

The current status of domestic research is as following:

(1) The corpus that is under construction in our country: it is mainly the media corpus, the student composition corpus, the spoken language corpus of the foreign students and so on;

(2) The subtitle is mainly focused on the translation research and the language characteristic research;

(3) In the aspect of application software, American blockbuster and classic TV series are generally used to assist in watching movies and learning English with subtitle tracking and control techniques.

The real Chinese film subtitles are used as the corpus, and there is no research literature and software to associate and excavate the Chinese characters grade outline, the vocabulary grade outline, the culture grade outline, the grammar grade outline. This can be either a entertainment field or a learning field, learning Chinese under the guidance of a teacher, or learning Chinese in a non-dominant situation.

III. SYSTEM DESIGN

Drawing on the experience of the construction of the Chinese media corpus, the students' written composition corpus and spoken language corpus, and the Chinese word segmentation system, the Chinese indexing system and the Chinese language statistics system, which are implemented on these corpora, The experience is applied to the subtitle corpus of Chinese movies and TV programs, and the subtitle dynamic corpus is built, and the subtitle corpus is linked with the Chinese character grade outline, the vocabulary grade outline, the cultural grade outline, and the grammar grade outline. To form a set of teaching programs for fun and fun to learn Chinese, and to create a software that students can both watch movies and learn Chinese (while watching Chinese movies, television programs, while learning Chinese). The software has both Chinese learning and entertainment functions. Also can realize the automatic test, the specific research content is as follows:

(1) Choose the appropriate subject matter, language standard film, television program as the multimedia material of software, and build the dynamic corpus with the Chinese subtitles of these materials;

(2) Translation and proofreading of subtitle data;

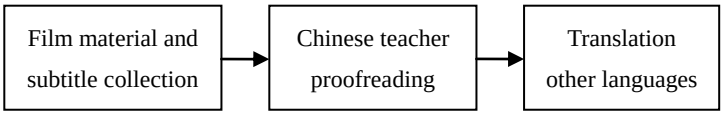

Fig. 1 Material finishing workflow

(3) Word segmentation, indexing, statistical research, as shown in Fig. 2;

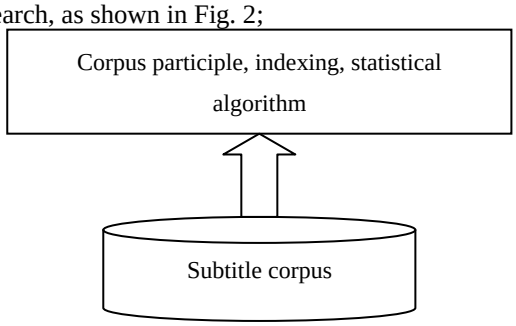

Fig. 2 Subtitle corpus processing

(4) The research on the association mining between subtitle corpus and Chinese character grade outline, vocabulary grade outline, culture grade outline and grammar grade outline.

(5) On the basis of the association mining with the grade outline, a data dictionary based on the dynamic corpus, which can be used for students' associative learning and associative learning is constructed, which is aimed at the existing lexicon, culture and grammar of the existing corpus. It is as a dictionary basis for automatic test.

(6) The main functions of the software are studied, including the control technology of the subtitle, the editing and tagging technology of the subtitle, etc. The main functions of the software are designed as follows:

① Teachers or students can mark words, culture, grammar, annotations, key sentences, etc.

② Subtitle control: display or hide the original

text, translation, annotations, new words, key sentences.

③ Learning resources, easy self-control: if a student on the network to see some video and audio materials used for learning is very good, then he can easily use software to make and learn.

④ The software provides recording and dictation functions, such as karaoke while listening, original or role-playing audio files can be exported, and can set some sentences in the subtitle reading times, so as to learn during the journey and ride. The second phase of the project can continue to develop the Android version of the software, so that students can also use the software on mobile devices.

⑤ Speaking practice, role-playing: it is a very effective means to improve spoken language by imitating the sounds of phonetic materials such as movies and TV plays. The student can record his voice and compare it to the original voice to find and correct his pronunciation problems. He can choose some characters in the film and television to play and then play the role of speaking.

(7) According to the language level, the research and implementation of the system automatically.

IV. TECHNOLOGY REALIZATION

(1) Resource Collection: classic Chinese movies and TV shows

Artificial collection, artificial subtitles, and invited Chinese teachers to screen those movies, television programs, songs suitable for making software multimedia materials.

(2) Subtitle translation

Chinese teachers can proofread and proofread the multimedia materials by subtitles, and for some languages, they can invite students who have a good command of the language (English, Indonesian, Thai, etc.) to do translation and proofreading.

(3) Construction of subtitle corpus

The subtitles of each film are a corpus, which is constructed by parallel corpus and dynamic corpus.

(4) Word segmentation and statistics of subtitle corpus

On the basis of the existing technology, it is the subtitle corpus of the word segmentation, statistics.

(5) Mining knowledge of Chinese characters, vocabulary, culture and gramma

In the light of the Chinese language teaching and the syllabus of the Chinese language teaching in the Chinese language institute, the supervised and unsupervised data mining classification algorithms are used to excavate the graded words, culture and grammar in the subtitles. For the machine which is difficult to realize, Chinese teachers need manual tagging, of course, the machine mining results also need Chinese teachers' manual proofreading.

(6) Software development and function design

Based on C++/DirectX/WTL/Windows SDK/XML/web service (gSoap+java) technology to develop learning software and automatic test system.

V. SUMMARY

(1) In this paper, the teaching of Chinese is combined with production, learning and research, which is the synergy and integration of different social division of labor in function and resource advantage in scientific research, education and production.

(2) The comprehensive application of software development, corpus and participle statistics provides service for Chinese teaching.

(3) Subtitle corpus and Chinese character rank outline, vocabulary grade outline, culture grade outline, grammar grade outline are associated mining and expert tagging, and the corresponding Chinese learning knowledge points of the film are generated.

(4) A more specialized teaching tool has been developed for interesting Chinese course: appreciation of Chinese movies and singing Learning Chinese.

(5) For interesting Chinese learning provides a one-stop learning and test solutions.

REFERENCE

[1] Zhang Jiaqi. A Corpus-based study of the language characteristics of American TV dramas in English subtitles: 2008(9)

[2] Jiang Lu. A Corpus-based study of the vocabulary characteristics of American TV dramas in English subtitles: 2012(5)

[3] Zhang Renxia. Corpus Retrieval of Black English in the subtitles of Akila and Scrabble Competition [J]. Film reviews: 2010(11)

[4] Zhang Renxia, Zhong Jian. Examples of back Translation of Chinese words by Bilingual Film subtitle Corpus and its implications for Translation Teaching [J]. Journal of Changchun University of Technology: 2012(6)

[5] Zheng Lilei, Xie Lei, et al. Design and implementation of automatic Chinese News subtitle Generation system [J]. Journal of Electronics: 2011(3)