Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

CYCCDC: A Chiayi Chinese Conversation Dialogue Corpus

Jui-Feng Yeh, Yun-Yun Lu, Yi-Syun Tan
Department of Computer Science and Information Engineering, National Chiayi University
No.300 Syuefu Rd., Chiayi City 60004, Taiwan (R.O.C.)
ralph@mail.ncyu.edu.tw, s1020447@mail.ncyu.edu.tw, s1000441@mail.ncyu.edu.tw

Abstract: Speech is one of the most natural ways of communication between human. In recent years, the spoken dialogue systems on human machine interaction (HMI) is more and more popular. In order to develop effective and natural human machine interaction, the corpus collected is relatively more important. Due to various types of corpus, classifying the corpus is a needed process. In this paper, we collected, transcribed, and classified the corpus, we named ChiaYi Chinese Conversation Dialogue Corpus (CYCCDC). We collected this corpus with multiple ways, and then we arranged and classified this corpus. The corpus includes multiple useful information to research spoken dialogue system and human communication field. The CYCCDC includes tourism information, food information, clothing, housing information, traffic information and part of Orange Technology, such as elderly health care and accident handling. This corpus can be extensively applied, such as tourism plan in spoken dialogue system.

Keywords: CYCCDC, Chinese corpus, tourism, orange technology, spontaneous speech.

1. Introduction

With the advance of the internet and technology, people chat with each other not only through voice. People can also make a communication by text with various well-developed online instant messaging software, such as Skype, Line, Aim, and so on. Since text is easier to record than voice, it is helpful for searching history record or finding out particular conversation information with proper classification in the future. There is some corpus analysis as (Agrawal, 2011) which is classified the emotions of the Hindi corpus. In (Jia et al., 2011), the present study systematically states the construction of the corpus on the English learners in Asia. There is also some conversation corpus, which has been collected in (Bechet et al., 2012), the goal of this paper is to reduce the development cost of speech analytics systems by reducing the need for manual annotation. In (Takezawa et al., 2002), they collected the travel conversation corpus and a broad-coverage bilingual basic expression corpus, and they compared the characteristics of vocabulary and expressions between these two corpus. In this paper, we collect the conversation transcript documents and analyze the topic classification. Finally, we gave each conversation script a topic classification. In particular, the corpus of care and accident handling are classified into at the category of Orange Technology (Wang & Chen, 2011). Orange technology is the idea mentioned by the National Cheng Kung University professor Jhing-Fa Wang. The main idea of orange technology is to bring health, happiness and care for human. It also include elderly health care and child care. Besides, people can enhance safety and quality of humanism to foreign culture with orange technology. We also collected other corpus on several topics: tourism, food, sports, solicitude and others. Each topics can be classified into more categories. Ohtake et al. proposed a new corpus of consulting dialogues which is designed for training a dialogue manager (Ohtake et al., 2010). They also collected more than 150 hours of tourist guidance dialogue. In section 2, we describe the method how we collect corpus. Then we explain how we arrange and classify in section 3. In section 4, we show that our corpus can be applied in various fields. Eventually, we discuss the future work and conclusions in section 5 and 6.

2. Method

In this paper, We collected the Chiayi Chinese conversation dialogue corpus to increase the variability of response sentence for the spoken dialogue systems. Our approach is to ask at least two people for a chat and record their conversation. There are several rules as following:

- Sentence composed by the conversation dialogue (unlimited number of the sentences)
- Period of each conversation occurred at least 8 minutes
- Each dialogue included at least two topics
- Each conversation took by native Chinese speakers

The rule 1 is established to make the system spoken dialogue more humane. The purpose of the rule 2 is to let speakers converse in a spontaneous way. The rule 3 is helpful to increase the variability of contents that can enrich response sentence. In order to reduce Chinese grammar errors and obtain a native Chinese spoken dialogue, we augmented the rule 4. We obtained 392 audio files which length is at least 8 minutes, then turned the audio files into the transcribed text files. There are at least 20 sentences in each transcribed file. The sentences are related to the fields of tourism information, elderly care and accident disposal, which included health, unexpected events, food, traffic, housing, clothing and others. An example of the conversation is shown in Figure 1.

Speaker A：這個周末我也要去嘉義玩
Speaker A：希望也能有好天氣。
Speaker B：你要去嘉義哪裡玩?
Speaker A：先去看射日塔之後再去阿里山看日出吧。
Speaker B：二二八紀念碑也可以去看看。
Speaker A：要怎麼去呢?
Speaker B：你可以坐公車去。
Speaker A：附近有加油站嗎?
Speaker A：我想開車去嘉義。
Speaker B：附近過橋有一家中國石油。
Speaker A：嘉義有甚麼好吃的在地小吃嗎?
Speaker B：有一間簡單火雞肉飯不錯吃。
……

Figure 1. An example of the conversation.

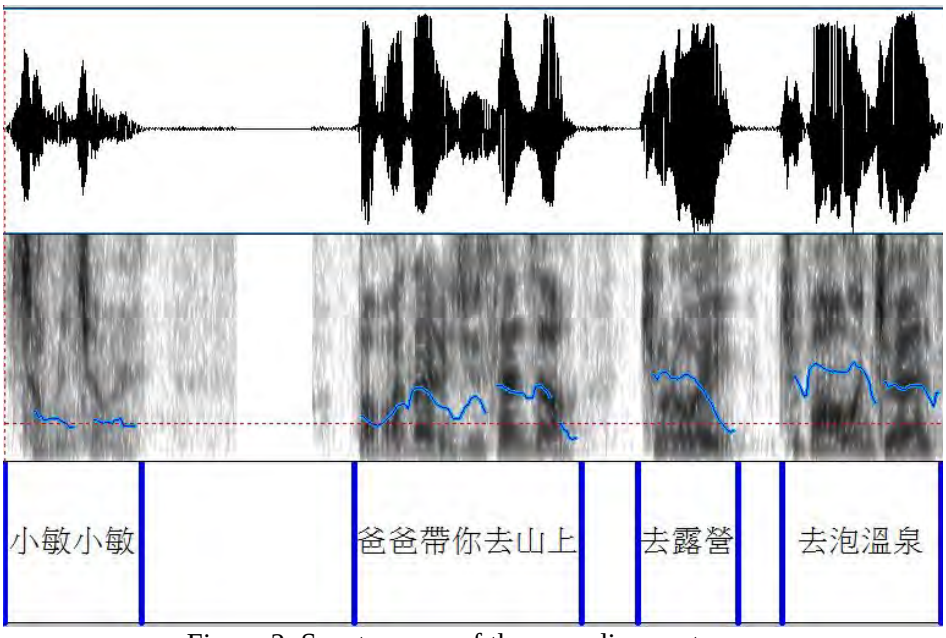

Figure 2. Spectrogram of the recording sentence.

First, the conversations must be recorded in a quiet room. Speakers need a computer, a microphone, and the recording software Praat before starting a conversation.. Audio file format is designated as the WAV format, with a bit rate of 256 kbps and a sampling rate of 16 kHz. The sampling resolution in 16 Bit is recorded in a mono mode with the PCM. The speakers can think about what they want to talk about before recording. Each conversation should be took at least 8 minutes and recorded at least 24 minutes in total. An example of Spectrogram is shown the show in figure 2.

The Chiayi Chinese conversation dialogue corpus has been recorded by 199 speakers, which are the students of National Chiayi University, Taiwan. There are 88 native Taiwan females and 111 native Taiwan males. The age range of these speakers is from 18 to 21. In total, there are 399 audio files and 27.5 hours of conversation speech. Consequentially, the Chiayi Chinese conversation dialogue corpus comprised 9,734 sentences of 399 transcribed text files.

3. Corpus Analysis

Table 1: Instruction of the topic.

Topics	Instruction
Tourism	Dining Location Lodging Location Modes of Transport (Living Travel) Navigation Information (Modes of Transport) Traveling Spot Shopping Information (Living Travel) Cost of Time Weather/Climate Care Agency (Location)
Food	Dinging Place (Restaurant、Snack Bar) Transportation (The Modes of Transport Go for Dining) Cost of Time (Meal Time)
Healthy Care	Transportation (Healthy Care) Navigation Information (Healthy Care) Shopping Information (Healthy Care) Weather/Climate (Health Care) Health Status
Accident Handle	Emergency Incident Detection (Fall Down, Faint, Asthma Attack)
Recreation	Leisure Entertainments (Knowledge/Reading, Recreational, Artistry, and so on)
Solicitude	Health State Greeting Consolation Family/Neighbor Trivia Working And so on
Sports	Ball Running Swimming Hiking And so on
Others	Not classified in the above topics

In this section, we describe how we design and classify these collected corpus. Due to a wide variety of corpus, we sorted all the corpus before the classification. It will make developers more convenient while using our corpus. And we will introduce main topics in section 3.1, then spoken speech phenomenon in section 3.2. In section 3.3 is describe about what is our method to distinguish topic in the sentences.

A “topic” is an abstract or a representative summary of the contents in dialogues. Developers can comprehend key points with the information of the topic instead of reading whole contents of dialogues. We can also classify conversations into appropriate categories. For example, a speaker says: “這個週末連假我們去嘉義的阿里山走走吧”, this sentence not only can be regarded as an effect of the topic "tourism" from a conversation, but also speculate from the keyword that the topic of this diaglogue might be “tourism”. If there exists a topic in one interpersonal conversation this conversation would not be interrupted easily. Besides, this uninterrupted conversation can be continued to develop new contents or changed into different topic timely can motivate speakers for keeping the interpersonal conversation. Therefore, we designed 8 topics to classify the corpus. There are tourism, food, healthy care, accident handle, recreation, solicitude, sports and others. Each topic can be subdivided into a lot of classes as shown in table 1. According to transcription which is tagged each topic by speakers. There are 2,064 sentences in tourism, 1,042 sentences in food, 177 sentences in healthy care, 268 sentences in accident handle, 1,801 sentences in recreation, 2,128 sentences in solicitude, 513 sentences in sports and 1,741 sentences in others. The distributing graph is shown in figure 3. There are several issues need to be discussed.

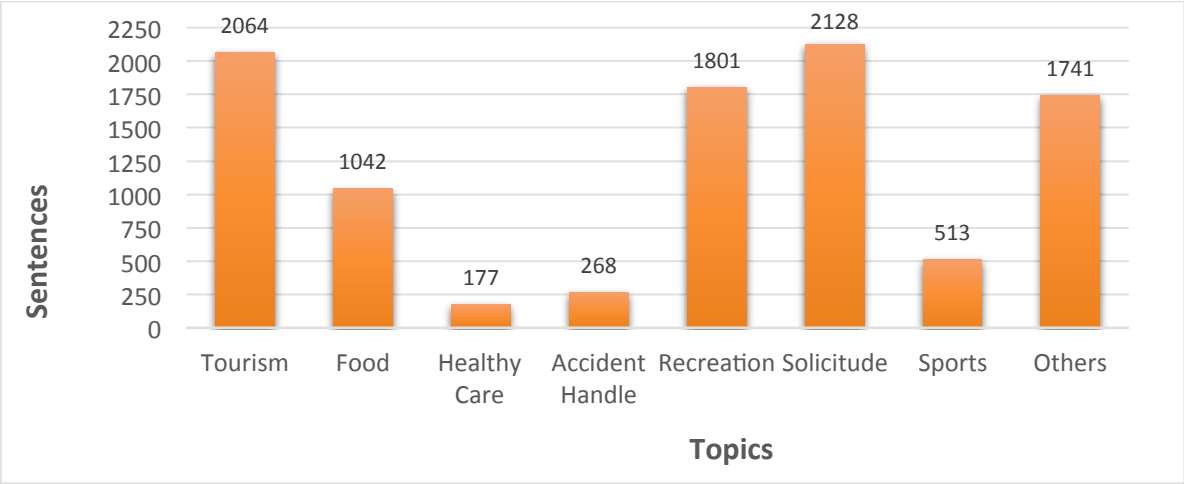

Figure 3. The distributing graph of each topics sentence.

3.1 Main Topics

According to human thought, the dialogue is nothing more than their food, clothing, housing, traffic and entertainment for main topic. For example, an old sick people want to travel, then he goes to some fun places, eat goodies and experience what he could not engage in leisure activities during illness. In this example, we designed the 7 topics, including the examples mentioned in the travel and food, while healthy Care, accident handle and solicitude are especially designed for the elderly.

3.2 Spoken speech Phenomenon

Through listening to the recorded speech, we found that corpus contains many voices speaking tone, because people thinking and emotion factors. Due to the recording of the conversation dialogue is not like reading speech which has a transcript. There is some spoken speech phenomenon. The most common of spoken speech phenomenon is particles, and the other is paralinguistic phenomena, the situation of pronunciation is not correct that is pronunciation error, and non-native language (Chang et al., 2005). There are an example of paralinguistic phenomena shown in figure 4. Because of the speakers conversed spontaneously, we classify those corpus to other topic.

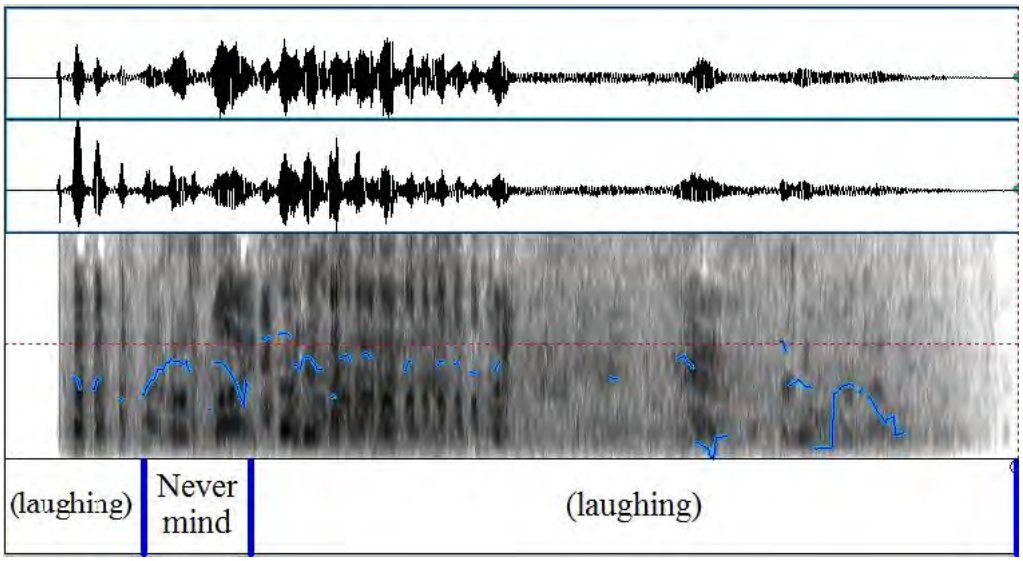

Figure 4. Spectrogram of the laughter in conversation (paralinguistic phenomena).

3.3 Keywords of topic

For example, a speaker A says: “最近有聽說什麼好吃的美食嗎?”, a speaker B says: “有阿!聽說在嘉義奮起湖有好吃的鐵路便當”. There has different keyword of topics in this example. We will judged according to the context of the sentence belongs to what topics.

4. Applications

The most application of corpus is automatic speech recognition (Wu et al., 2014). There is also another application of corpus, such as generating responses sentences. Since the CYCCDC contains many different topics of sentences and rich vocabularies, the CYCCDC can applied in many aspects. For instance, developers can utilize the CYCCDC to generate responses sentences with some tour information for a tourism planning system, or refine an existing dialogue system with corpus of solicitude. The developers also can exploit the CYCCDC to analyze what do people talk about or when they want to have a travel. The CYCCDC is a conversational corpus so that the generated responses sentences of a dialogue system can make users feel like having a conversation with an actual human. This type of dialogue system is called as chat oriented dialogue system. Banchs et al. have also proposed the informal response interactive system (IRIS) (Banchs et al., 2012), which is a chat-oriented dialogue system based on the vector space model framework.

5. Future Work

In the future work, we will continue to improve the consistency of the database. With more corpus from the refined CYCCDC, all the related applications can be expected to be improved greatly, and can be applied in more studies. The spoken speech phenomenon also is still a research issue.

6. Conclusion

In this paper, our corpus was recorded and transcribed by the speakers. Different from other common corpus, the Chiayi Chinese conversation dialogue corpus is based on conversation for academic research but also substantial contribution. We designed 8 topics of the CYCCDC, which included the vast majority of tourism and solicitude. The native regional range of CYCCDC involved Yunlin, Chiayi and Tainan in Taiwan. Thus, the corpus also can apply in tourism planning system which

focused on these regions. Although all the sentences are still in the testing phase, but the quality of the sentences is adequate enough for doing researches. The other parts of corpus in CYCCDC are mainly collected for the Orange Technology, which focused on health care and accident handling for children and elders. The CYCCDC also involoved some dialogue about disadvantaged groups or people who need social care. This part of the corpus can assists with the academic research of society. Certainly, the CYCCDC also contained some of common corpus, such as recreation, sports and other categories that are irrelevant to the main collection. Because the CYCCDC is a real conversation between humans, the CYCCDC is also helpful for a chat-oriented dialogue system.

Acknowledgments

This work is supported in part by the National Science Council, Taiwan, R.O.C., under the project grant numbers NSC 102-2221-E-415-006-MY3.

References

Agrawal, S. S. (2011, October). Emotions in Hindi speech-analysis, perception and recognition. In Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on pp. 7-13. IEEE.

Jia, Y., Wang, M., Zhai, H., & Li, A. (2011, October). Construction of speech corpus of AESOP-SD. In Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on pp. 42-46. IEEE.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Beze, M., De Mori, R., & Arbillot, E. (2012). DECODA: a call-centre human-human spoken conversation corpus. In LREC on pp. 1343-1347.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., & Yamamoto, S. (2002, May). Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In LREC on pp. 147-152.

Wang, J. F., & Chen, B. W. (2011). Orange computing: challenges and opportunities for awareness science and technology. In 2011 3rd International Conference on Awareness Science and Technology (iCAST) pp. 533-535.

Ohtake, K., Misu, T., Hori, C., Kashioka, H., & Nakamura, S. (2010). ”Dialogue acts annotation for NICT Kyoto tour dialogue corpus to construct statistical dialogue systems” Seventh International Conference on Language Resources and Evaluation, LREC 2014, pp. 2123-2130.

Wu, C., Shen, H., & Yang, Y. (2014) “Chinese-English Phone Set Construction for Code-Switching ASR Using Acoustic and DNN-Extracted Articulatory Features” IEEE/ACM Transactions on Audio, Speech & Language Processing 22(4), IEEE, 2014, pp. 858-862.

Banchs, R. E., & Li, H. (2012). “IRIS: a Chat-oriented Dialogue System based on the Vector Space Model” Proceedings of the Association for Computational Linguistics 2012 System Demonstrations, pp. 37-42

Chang, L. H., Wang, Y. R., & Chen, S. F. (2005) “Evaluation of Mandarin Broadcast News Transcription System” ROCLING 2005