Building Large Chinese Corpus for Spoken Dialogue Research in Specific Domains

Changliang Li, Xiuying Wang
Institute of Automation, Chinese Academy of Sciences
{changliang.li,xiuying.wang}@ia.ac.cn

Abstract

Corpus is a valuable resource for information retrieval and data-driven natural language processing systems, especially for spoken dialogue research in specific domains. However, there is little non-English corpora, particular for ones in Chinese. Spoken by the nation with the largest population in the world, Chinese become increasingly prevalent and popular among millions of people worldwide. In this paper, we build a large-scale and high-quality Chinese corpus, called CSDC (Chinese Spoken Dialogue Corpus). It contains five domains and more than 140 thousand dialogues in all. Each sentence in this corpus is annotated with slot information additionally compared to other corpora. To our best knowledge, this is the largest Chinese spoken dialogue corpus, as well as the first one with slot information. With this corpus, we proposed a method and did a well-designed experiment. The indicative result is reported at last.

1 Introduction

Spoken dialogue system is regarded as the original form of the famous Turing test, as well as a long goal in artificial intelligence and natural language processing field. Though much progress has achieved, the research for spoken dialogue in specific domains faces many challenges, such as query understanding and the response generation and so on. Large-scale conversation corpus with rich annotation will offer great support to the research.

With the development of network and communication technology, there is a vast amount of data available documenting human communication. Much of them could be used, perhaps after some preprocessing, to train a dialogue system. English receives much more research attention than any other languages. In the last few decades, several English conversation corpora have been published, such as Carnegie Mellon Communicator Corpus (Bennett and Rudnicky,2002), Dialog State Tracking Challenge (DSTC) (Williams et al.,2013) and DIALOG mathematical proof dataset (Wolska et al., 2004). However, little research on other language discourse have been reported, which limits the growing research interest in these languages.

In this paper, we mainly focus on Chinese corpus creation for its wide range of use and its popularity. To overcome this problem, we build a large-scale corpus of spoken dialogue system, which consists of five domains and more than 140 thousand dialogues. To our best knowledge, it is the largest scale Chinese spoken dialogue corpus. Moreover, different to all other corpora including ones in English and any other languages, each sentence in our corpus is annotated with rich slot information. It is the first corpus annotated with this kind of information, which will play a major role in building spoken dialogue system.

The structure of the paper is as follows. Section 2 introduces some related works on spoken dialogue corpus. Section 3 lays out the process of corpus construction and the statistics about our corpus. Section 4 proposes a method combination

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 320–324,
Taipei, Taiwan, November 27 – December 1, 2017

with the constructed corpus and presents the result. Section 5 gives the conclusions to end the paper.

2 Related word

In this section, we review some English corpus and some related efforts for building Chinese corpus. Since there are many English corpora, we just list some examples.

1) Switchboard dataset (Godfrey et al.,1992)

One popular corpus is Switchboard dataset (Godfrey et al.,1992). It consists of approximately 2,500 dialogues over the phone from 500 speakers, along with word-by-word transcriptions. About 70 casual topics were provided, of which about 50 were frequently used. The corpus has been used for a wide variety of other tasks, including the modeling of dialogue acts such as 'statement', 'question', and 'agreement' (Stolcke et al., 2000).

2) The Ritel corpus (Rosset and Petel, 2006)

It is a small dataset of 528 spoken questions and answers in a conversational format. The purpose of the project was to integrate spoken language dialogue systems with open domain information retrieval systems, with the end goal of allowing humans to ask general questions and iteratively refine their search. The questions in the corpus mostly revolve around politics and the economy, along with some conversations about arts and science-related topics.

3) The DIALOG mathematical proof dataset (Wolska et al., 2004)

It is a Wizard-of-Oz dataset involving an automated tutoring system that attempts to advise students on proving mathematical theorems. The system is completed by using a hinting algorithm that provides clues when students come up with an incorrect answer. At only 66 dialogues, the dataset is tiny and consists of a conglomeration of text-based interactions with the system, as well as think-aloud audio and video footage recorded by the users as they interacted with the system.

Chinese corpus is few, and we list two relatively popular ones.

4) CASIA-CASSIL(Zhou)

It is a large-scale corpus of Chinese casual telephone conversations in tourism domain. The source data is collected from a large number of spontaneous telephone recordings up to the present. After a strict selection, only a minority of dialogs remains, which are with good voice quality, enough turns and strictly belong to required domains.

5) Lancaster Los Angeles Spoken Chinese Corpus (LLSCC)

It is a corpus of spoken Mandarin Chinese. The corpus is composed of 1,002,151 words of dialogues and monologs, both spontaneous and scripted, in 73,976 sentences and 49,670 utterance units (paragraphs). LLSCC has seven sub-corpora, Conversations: Telephone Calls Play & Movie Transcripts, TV Talk Show Transcripts: Debate Transcripts Oral Narratives and Edited Oral Narratives.

Both Chinese corpora cannot fulfill the demand for Chinese spoken dialogue system research. So we built a larger and richer corpus in this paper.

3 Corpus Creation

In this section, we describe the steps we took in the construction of our corpus. Figure 1 gives an overview of the main workflow of our corpus creation. We will illustrate the details in the following subsections.

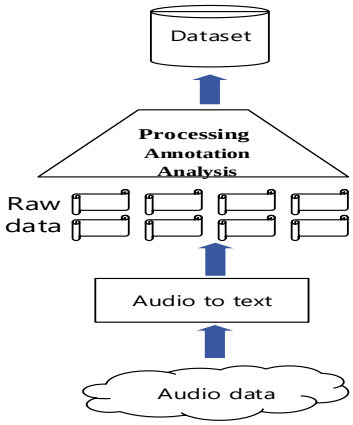

Figure 1 Overview of the main workflow of our corpus creation

3.1 Collection

The source of the corpus is very crucial to determine the quality of corpus. The dialogue data in a specific domain in real life is often combined with commercial secrets as well as with peoples' private information. For example, when a person books ticket through phone, he/she has to offer private information. So it is hard to obtain such data from a website or by some free methods. That is why there is a little corresponding corpus of dialogue in particular domains.

While collecting original data, we are assisted by volunteers, who offer the data from their real life for research purpose. The data consists of their audio data (phone recorders) while they call for some service, such as booking restaurant or hotel.

Since we have got the original data, we then employ speech recognition technology to transfer audio to text.

3.2 Data Processing and Annotation

Since we have the real data, we make two steps following. Firstly, for protecting the privacy right, we remove the private information. We used the encoded token to replace any information which may release private information such as name, phone number and so on.

Second, after analyzing the data, we find that there is much rough data, due to either the error of transferring audio to text or meaningless sentences caused by the bad habits of oral communication. Furthermore, many people are hired as expert assisted roles to select high-quality data and make the text more formal.

Slot information is required in specific domains. We manually extracted the slot as attachment information of dialogues. The slot number is limited in one specific area, for example in the domain of weather querying there are only two slots, time and location. Table 1 gives the example of the final version in booking hotel domain.

q: 您要什么房间，要几间？ {What type room will you book, and how many rooms do you want?} [0] [0] [0] [0] [0] [0]
a: 我想订三间总统套间。 { I'd like to reserve three presidential suites } [0] [0] [0] [0] [0] [0]
q: 住几天您？ { How long are you going to stay? }
[总统套间] [三] [0] [0] [0] [0]
a: 就住一天就行。 { one day is ok}
[总统套间] [三] [0] [0] [0] [0]
q: 您贵姓啊？ {would you like to offer you name ,sir?}
[总统套间] [三] [0] [一天] [0] [0]
U: 我的称呼是吴佳。 {My name is wujia}
[总统套间] [三] [0] [一天] [0] [0]
M: 什么时候来？ { When would you like to check in? }
[总统套间] [三] [0] [一天] [吴佳] [0]
U: 大概五月八号下午五点吧。 { At about five p.m. on May 8th }
[总统套间] [三] [0] [一天] [吴佳] [0]
M: 你给我您的电话。{ sir, please tell my your phone number }
[总统套间] [三] [五月八号下午五点] [一天] [吴佳] [0]
U: 联系号码是 012321322。 {it's 012321322 }
[总统套间] [三] [五月八号下午五点] [一天] [吴佳] [0]
M: 好的，吴佳，为您预订了五月八号下午五点总统套间三间。{ ok, sir, we reserve three presidential suites for you at five p.m. on May 8 }
[总统套间] [三] [五月八号下午五点] [一天] [吴佳] [012321322]

Table 1: dialogue example in booking hotel domain.

The content in braces is the corresponding English version for better understanding. The content in brackets is corresponding slots information. 0 means the slot is empty at present. In this example, the slots are room type, number of rooms, the data check in, stay time, customer name and phone number.

3.3 Analysis

As a result, we have got the final version corpus of 5 domains. The detailed statistics are summarized in this subsection.
Table 2 gives the slot distribution in each domain.

Domain	slot	num
Booking restaurant	Time/number of people/name	3
Booking hotel	Room type/number of rooms/time to check in/stay time/name/phone number	6
Weather query	Time/location	2
Ordering taxi	Destination/departure/name/phone number	5
Top up	Phone number/ Amount of money	2

Table 2: slot distribution in each domain

We can see that booking hotel domain has the most slots. Weather query and top up has only two slots.

Figure 2 gives the dialogue distribution in each domain.

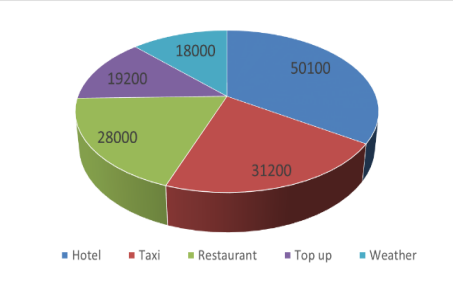

Figure 2: The dialogue distribution in each domain

From Figure 2, we know that booking hotel domain has the most dialogues, and weather query domain has the least dialogues. This distribution relates closely to the slot distribution.

Figure 3 gives the sentence length distribution of each domain.

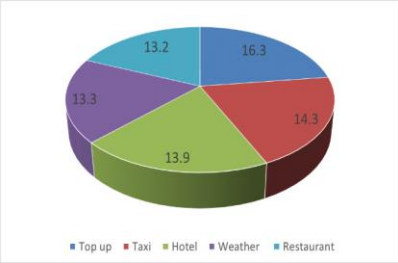

Figure 3: The sentence length distribution

From Figure 3, we know that the average length of sentences in each domain is similar.

Figure 4 gives the Dialogue turn distribution in each domain.

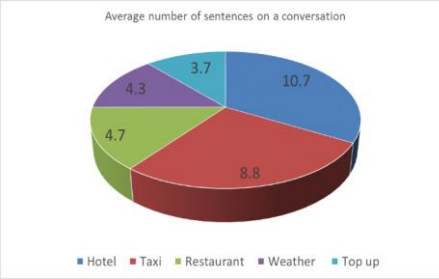

Figure 4: The Dialogue turn distribution

From Figure 4, we can see booking hotel and ordering taxi needs more turns to finish the dialogue. This distribution relates closely to the slot distribution. It is easy to understand that more slots, more turns required to obtain all information.

Figure 5 gives the word distribution in the sentences.

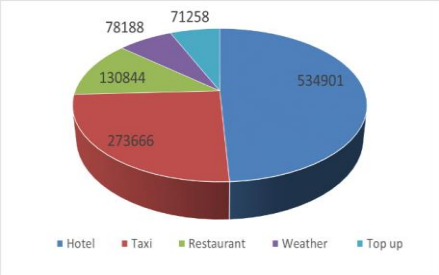

Figure 5: The word distribution in sentence

From Figure 5, we know that average length of sentences in each domain is similar.

From the statistics, we can see that the distribution of our corpus is relatively stable.

4 Method Based on the Corpus

Based on the corpus, we develop a model for spoken dialogue research in specific domains. The basic idea of the method is to introduce slot information in the corpus into the sequence-to-sequence framework.

After a query is given, the first slot is set as zero as there is no slot information at the beginning. Slot information is extracted at next turn, so the slot state is changed by fulfilling the corresponding information. The dialogue system works gradually like this and as a result, finish the dialogue when all the slot information received.

Evaluation metrics for dialogue is still an open problem. At present, the popular metric is the turns that the model needs to fulfill all the slots and finish the process. The experiment result shows that combination with the created corpus, it requires almost the same turn as used in real life. However, there are still many open problems worthy researching.

5 Conclusion

In this paper, we introduce a high-quality and large corpus for spoken dialogue research in specific domains. The corpus consists of five different domains and more than 140 thousand dialogues. All the data are created based on real life data. As our best knowledge, this is the largest Chinese spoken dialogue corpus, as well as the first one with rich slot information. We believe that the corpus will greatly support the spoken dialogue system research.

References

C. Bennett and A. I Rudnicky. 2002.The Carnegie Mellon communicator corpus.

J. Williams, A. Raux, D. Ramachandran, and A. Black. 2013. The dialog state tracking challenge. In Special

Interest Group on Discourse and Dialogue (SIGDIAL).

M. Wolska, Q. B. Vo, D. Tsovaltzi, I. Kruijff-Korbayová, E. Karagjosova, H. Horacek, A. Fiedler, and C. 2004. Benzmüller. Anannotated corpus of tutorial dialogs on mathematical theorem proving. In The International Conference on LanguageResources and Evaluation (LREC).

Godfrey J J, Holliman E C, McDaniel J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Acoustics, Speech, and

Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. IEEE, 1992, 1: 517-520.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics(2000), 26(3):339–373.

S. Rosset and S. Petel. 2006. The ritel corpus-an annotated human-machine open-domain question answering spoken dialog corpus. In The International Conference on Language Resources and Evaluation (LREC).

Zhou K, Li A, Zong C. Dialogue-Act Analysis with a Conversational Telephone Speech Corpus Recorded in Real Scenarios.