A New Chinese Dialogue Corpus Extracted from Baidu Tieba

Akash Singh*, Alan An*, Miles Q. Li*
260728046, 260683828, 260786295
School of Computer Science, McGill University
{$akash.singh, chengyun.an, qi.li7$}@mail.mcgill.ca
(*equal contribution)

***Abstract*—Most dialogue corpora existing are in English. Therefore, we introduce a Chinese dialogue corpus designed for training goal-driven end-to-end dialogue systems.**

I. Introduction

We introduce a Chinese corpus suitable for goal based data-driven learning of dialogue systems. This corpus is constructed using human-human conversations on topics related to economics and economy. The corpus is web-crawled over `https://tieba.baidu.com/` and is available with the source code at `https://bitbucket.org/comp551proj1/proj1` .

II. Dataset description

Text corpora have always been used by philosophy, historical linguists, and lexicographers; but over the past decades, the availability of large electronic corpora have significantly extended the possible research for natural language processing tasks including dialogue systems [1]. Before we explore those interesting possibilities, one major limitation that arises when considering the use of corpora is that existing corpora are mainly in English; for example, the top 10 corpora in Linguistic Data Consortium (LDC, `http://www.ldc.upenn.edu`) are all English corpora. With more than 6000 living languages [2], it is important that we analyze the possibilities towards corpora of different languages.

Chinese as a language is written without inter-word spaces, therefore its application in creating dialogue systems is more interesting. Dialogue systems for English has been an active area of research in computational linguistics for almost two decades and is a topic of active research around the world. No proper Chinese dialogue system has yet examined the handling of human-machine communication where there is little surrounding context. That is partly because the lack of large-scale multi-turn Chinese dialogue corpus. According to the official wiki of Baidu Tieba[1], it is the largest Chinese forum around the globe and it has 8.2 million sub-forums (topic specific), 3.5 billion posts, 64.6 billion replies in total. There is great potential to create a large-scale dialogue corpus with this forum. Therefore, in this report we propose an approach to create a Chinese dialogue corpus from Baidu Tieba. With this approach, anyone can create a Chinese dialogue corpus utilizing all the posts and replies in Baidu Tieba, the corpus we created focuses on dialogues about economics and economy though.

A. Type of Corpus

Forum corpora comprise of conversations on forum-based websites such as Reddit[2] or Baidu Tieba[3] where users can publish posts, and other users can make comments or replies on top of that post. In some scenarios, replies are nested indefinitely, as users make replies to previous replies. There is no restriction on the number of words of a post. Also there is no word limit like those on micro-blogging websites. Therefore, utterances in a forum-based corpora are usually long and the number of utterance in a dialogue is ideal [1].

The advantage of forum corpora is that they provide more generalized dialogues in reference to a goal. The human interlocutors are not aware that their posts will be used for corpus creation, so the dataset as a result is more natural [1]. The biases of gender, age and cultural background are removed in most popular forums. This also features a common characteristic which we want to simulate in our dialogue: $non-contactvarietiesofcommunication.$

B. Goal selection for the corpus

It is important to know that even the largest corpora can only represent a finite subset of a language's infinite potential. And given Zipf's law [3] that the frequency of use of the n-th most frequently used word in a corpus is inversely proportional to n, this way even the largest corpus will appear small for research. With the limitation of time and computational machines, we cannot create a very large corpus although the proposed method can technically create a corpus with at least millions of dialogues from Baidu Tieba considering its scale. In consequence, to alleviate such an issue, it helps to impose restrictions on the topics of the posts we will use since the vocabulary will be compact which is good for training a dialogue system. We looked into posts in several

[1]https://baike.baidu.com/item/%E8%B4%B4%E5%90%A7?fromtitle=%E7%99%BE%E5%BA%A6%E8%B4%B4%E5%90%A7&fromid=95221

[2]https://www.reddit.com/

[3]https://tieba.baidu.com/index.html

topics including science, economics, economy, global news, arts, calligraphy, public health etc. Some of them contain too little replies (e.g., science), some of them contain too many dirty talk (e.g., global news), some of them contain many talks centered on some image (e.g., arts and calligraphy) which would also be problematic because we do not wish to save the images and the dialogues would not make sense at all without being connected to the images. Others contain too many advertisements (e.g., public health). Considering all those factors, we decided to use topics related to "economics" or "economy" as the subject of dialogue corpus. According to our observation, posts related to economics and economy have a more educated audience and therefore we have a better chance of generating a high-quality corpus. Economy tightly connects every aspect of daily life together. Not only economy itself is discussed on economy-related forums, Social issues such as housing price, education reform are also hot topics on these forums. In the words of Lionel Robbins, "Economics is the study of given ends and scarce means." [4] Therefore it is a suitable topic not just for the quality of data, but also aiming for the quantity of data for corpora creation.

C. Data extraction and parsing

The recent eruption of number of people having informal communication on social media websites provides us a unique opportunity to get naturally occurring conversations that are exponentially larger than those previously available [5]. These corpora, in turn, present new opportunities to data driven dialogue systems. Data showing some theoretically interesting patterns can be obtained by web crawling of posts available on these websites without violating their copyright policies. So, we chose Baidu Tieba (copyright norms: `http://static.tieba.baidu.com/tb/eula.html`) to obtain our corpus. All the URLs related to economics or economy were crawled via python script using HTML and CSS selectors. A typical baidu URL resembles the following pattern-

```
https://tieba.baidu.com/p/5337319614
```

where "5337319614" is the "topic id". And to get the replies of a post under a topic, we need a URL like this-

```
http://tieba.baidu.com/p/comment?tid=5337319614&
pid=112532474244l&pn=1&t=1505875331044
```

where tid represents the "topic id", pid is the "post id", pn represents "page number" of the replies (only a post with more than 10 replies will be assigned a post id) and t is the "epoch time" in milliseconds. We then used all the crawled URLs to access the corresponding HTML and extract the posts and their replies. We parse users' posts by keeping a track of the HTML's tag value, #id and attributions associated with it such as the class. This was done using $Javajsoup$[4] library. The above approach can be generalized to crawl all the sub forums in Baidu Tieba and is not restricted to economics and economy forum. We tracked the mentioned parameters in a URL to extract this proposed corpus. Our corpus was collected by crawling exactly 929,346 posts across 56,997 web pages `https://bitbucket.org/comp551proj1/proj1/src/3788788f4753/URLcode/?at=master` specific to economics and economy.

Our data parsing program ensures the following consistencies:

1) *Semantic relevance* is the user content being correctly mapped to dialogues in the corpus. This requires the entities in the replies to be correctly aligned with those in the post. To ensure that, we use an temporary attribute "uname" to keep the user name for manually checking whether the mapping is correct. We eventually deleted these "uname" data along with this temporary attribute to ensure anonymity.
2) *Logical consistency* requires the content of the response to be logically consistent with the post. This consistency is checked by us and it is maintained in the generated corpus.

The dialog contains 83,719 utterances in 4,367 conversations. It is a research challenge to perform end-to-end learning from raw features to make high order decisions. One of the inter-disciplinary challenge involves the right formatting of the input dataset. Therefore, we formatted our corpus to best suite the requirement in designing an end-to-end dialogue system. In the proposed corpus, every conversation starts with a symbol <s> and an end symbol as </s>. Each utterance in the conversation is wrapped by the symbol <utt> and </utt>. Each human interlocutor is anonymized and is identified with a "uid" attribute of the <utt> token. The full document is enclosed between two tags <dialog> and </dialog>.

D. Data cleaning

We only keep those posts which have more than 10 replies in our corpus for the following reasons:

1) According to our observation short dialogues tend to be less interesting.
2) More advertisements and irrelevant replies exist in short dialogues.
3) With more turns, the dialogue will be much more useful to train multi-turn dialogue systems.

In addition, we removed most offensive utterances from our corpus utilizing the adapted version of Chinese short message service banned words[5]. To adapt this keyword list to our task, we recorded how many utterances were deleted because of each keyword and manually deleted those keywords from the list that did not affect the quality of the corpus. This was done to ensure that the dialogue system produced is of high quality without much loss. According to our observation, the deleted utterances are mostly very offensive or advertisements. We don't show them in this paper to avoid any discomfort but you can find the deletion log in our repository. After cleaning the dataset with the offensive keywords, some dialogues have less

[4]https://jsoup.org/

[5]https://github.com/spetacular/bannedwords

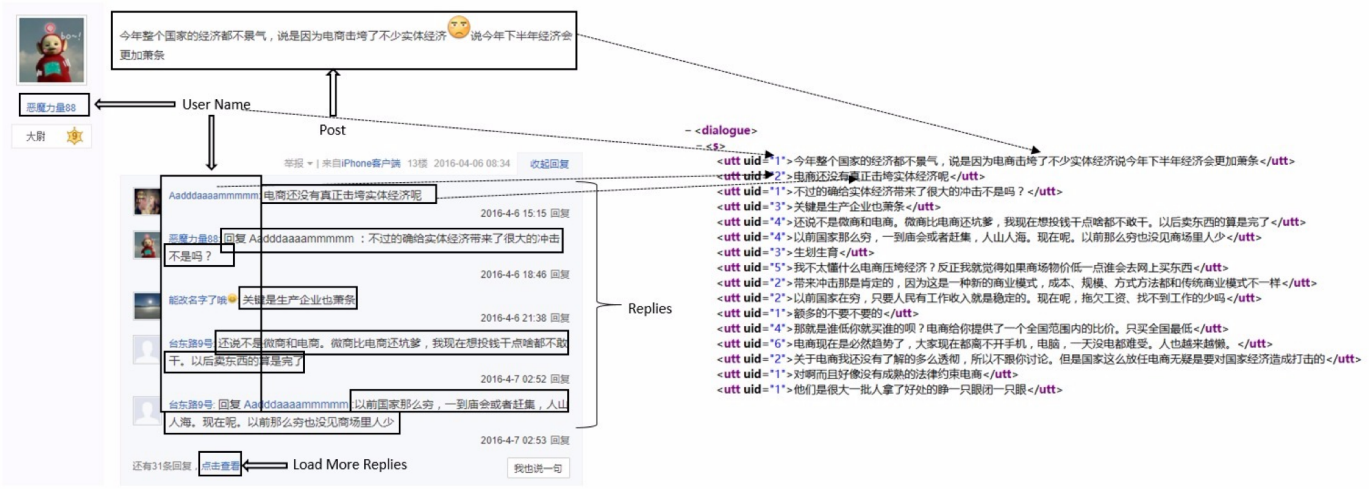

Fig. 1. Semantic Relevance and logical consistency of the corpus. An example of Baidu Tieba post and the replies it received and its mapping to dialogue in the corpus.

than 10 utterances. Final link to the cleaned data is mentioned in the footer[6].

E. Data statistics

The statistical data of the dialogue corpus is in Table I. It can be seen that an average 4 people (median is 3 people) are involved in a dialogue and their average utterance turn is 19. This is shown in Figure 2 and Figure 3.

Property	Value
Description	Posts and replies extracted from Baidu Tieba
Language (ISO639-2)	chi (B)
# of conversations	4,367
# of turns	83,719
# of words	3,820,251
# of utterances filtered for cleaning	2,691
Aveg.#of turns per conversation	19.2
Aveg.# of words per conversation	875
Aveg.# of words per turn	45.6

TABLE I
STATISTICAL DATA OF THE DIALOGUE CORPUS

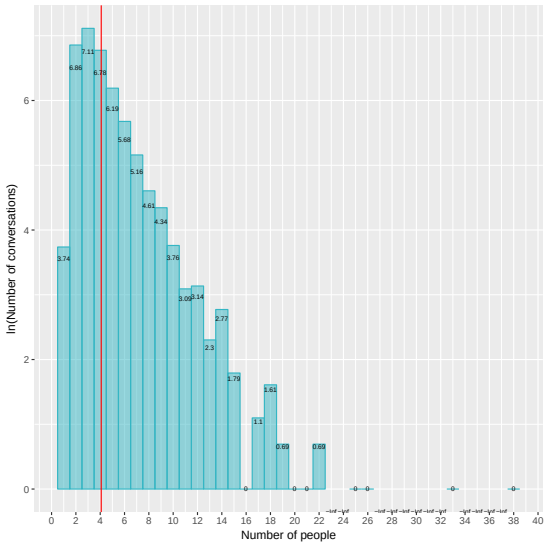

Fig. 2. ln of number of conversations plotted against number of people. It can be seen that an average 4 people (median is 3 people) are involved in a dialogue.

F. Data evaluation

We evaluate our corpus by comparing the number of non-Chinese words against the total number of Chinese words in our corpus. We found that out of 3,820,251 words we obtained a total of 13,202 non-Chinese characters. This corresponds to 0.35% of the text. When Chinese people talk with each other, they will use some English initials or proper nouns in their sentences, so this shows that our corpus is actually in Chinese. It can be seen that an average 4 people (median is 3 people) are involved in a dialogue and their average utterance turn is 19. This is reflected in Fig. 2 and Fig. 3.

III. DISCUSSION

According to our survey, there are some existing Chinese corpora and some of them are listed in this page: `http://github.com/candlewill/Dialog_Corpus`. Those corpora can be divided into two categories: speech-based and text-based. PolyU Corpus of Spoken Chinese proposed by Department of English, Hong Kong Polytechnic University[7] is a speech corpus. Although the authors have written transcripts of the speech, the texts still have the property of spoken language instead of written language. Besides, CADCC-Chinese Annotated Dialogue and

[6]`https://bitbucket.org/comp551proj1/proj1/src/f46d855fe4196e94614623cda068067cd882e705/Corpora/?at=master`

[7]http://www.engl.polyu.edu.hk/research/corpus/corpus.htm

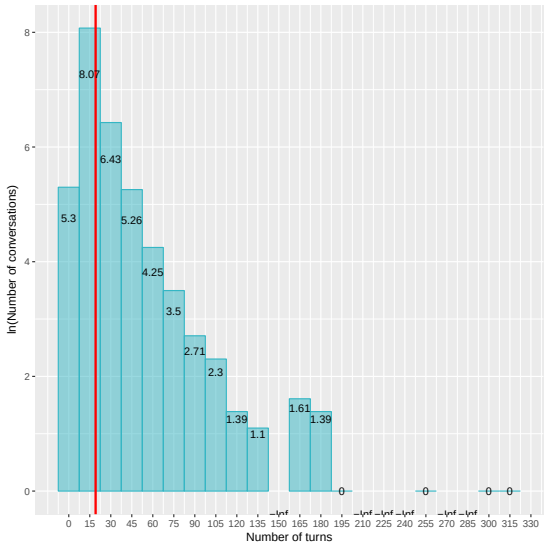

Fig. 3. ln of number of conversations plotted against number of turns.

Conversation Corpus[8] also originates from spoken language. So our corpus has different properties with theirs. As for text-based corpora, some are Question Answering Corpus including Insurance QA Corpus[9] [6] and Egret Wenda Corpus[10]. QA pairs can be seen as two-turn dialogues. Differently, our corpus are made up of multi-turn dialogues which is more powerful to train dialogue systems that are designed to have internal states to record the chatting history. The Chinese Conversation Corpus[11] is extracted from Chinese movie transcripts. Although this kind of conversations contain multiple turns, they still have the issue that the dialogues are not totally natural since they are written by scriptwriters instead of real life dialogues. Message Service Corpus contains 31465 real life messages but they are not organized in dialogues. They are separate messages so that they cannot be used directly to train dialogue systems. The existing corpus most similar to ours is the corpus described in Wang et al. [7]. They extract the dialogues from Sina Weibo[12] (a Chinese microblog website) [7]. Their corpus contains 46,345 posts and 1,534,874 responses which is larger than our corpus but as a giant corporation, they have many more computation resources and it takes them 2 months to extract the dataset. So it's not fair to compare the size of our corpus with theirs. Technically, their method can be used to extract all Sina Weibo posts and responses and our approach can extract all posts and replies of Baidu Tieba given enough machines and time.

IV. Application

In this report we propose a corpus for goal-based end-to-end dialogue systems. The model is based on medium-text based dialogue, to leverage the massive instances collected from Baidu Tieba. For research in similar directions, we create a dataset based on the posts and comments from Baidu Tieba. This dataset can be used for both training and testing automatic response models for short texts.

V. Statement of Contributions

All three group members together discussed about from what kind of sources to get the corpus and what the language of the corpus should be. We agreed that dialogues from books or transcripts were not natural since they were written by a writer. Considering Alan and Miles were both Chinese and it was convenient to clean the corpus if more group members were familiar with the language, we chose Chinese as the target language. That basically decided the subsequent division of job. Alan and Miles read the copyright norms of several candidate websites and made the decision to choose Baidu Tieba. Then all three of us decided the topic to crawl together. Akash crawled the pages contain economics and economy related posts and then Miles parsed the pages to extract the dialogues. Alan wrote the bash scripts to ensure the data was saved in the required form. In this process, we couldn't directly get all the responses to a post and Alan found a way to address this problem. Then Alan and Miles work together on reading the extracted dialogues and the post-processing including creating a keyword list to filter dirty talks and advertisements. Alan found the keyword list and Miles amended the list according to the observation of the filtered utterances. Then Alan wrote the Java methods and R script to get the statistics of the corpus and produced the graphs. Miles is in charge of merging the code. Akash made the draft of this report. Miles compared this corpus with existing corpora and drafted the Discussion section. We hereby state that all the work presented in this report is that of the authors.

[8]http://shachi.org/resources/24
[9]https://github.com/shuzi/insuranceQA
[10]https://github.com/Samurais/egret-wenda-corpus
[11]https://github.com/majoressense/dgk_lost_conv
[12]https://weibo.com

References

[1] I. V. Serban, R. Lowe, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," *arXiv preprint arXiv:1512.05742*, 2015.
[2] M. K. Anke Ldeling, *Corpus Linguistics*. Monton de Gruyter, 1997, vol. 2, pp. 297.
[3] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.
[4] L. Robbins, *An essay on the nature and significance of economic science*. Ludwig von Mises Institute, 2007.
[5] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 583–593.
[6] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: A study and an open task," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 813–820.
[7] H. Wang, Z. Lu, H. Li, and E. Chen, "A dataset for research on short-text conversations." in *EMNLP*, 2013, pp. 935–945.