Title	マイクロブログからの対話コーパスの自動構築
Author(s)	関田，崇宏
Citation	
Issue Date	2020-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/16408
Rights	
Description	Supervisor:白井　清昭，先端科学技術研究科，修士(情報科学)

Construction of Dialog Corpus from Microblog

1810104 Takahiro Sekita

In recent years, many studies of dialog systems that can chat with users are widely investigated. A dialog corpus, which is a collection of dialogs between humans, is necessary to develop such dialog systems. However, it is rather difficult to construct a large scale dialog corpus, since it requires much cost to record and transcribe conversation between human. On the other hand, several researchers attempted to automatically construct a dialog corpus by retrieving a large amount of sequence of tweets and replies, which is regarded as a pseudo dialog, from Twitter. However, such sequence of tweets and replies may not be a real dialog. One of the problems of the previous studies of automatic construction of a dialog corpus from Twitter was that they did not consider whether the retrieved pseudo dialogs were appropriate to be included in a dialog corpus. The goal of this thesis is to automatically construct a high quality and large scale dialog corpus by retrieving dialogs, i.e. sequence of tweets and replies, from a microblog (Twitter) and removing inappropriate ones from them.

The proposed method consists of three steps: collecting sequence of tweets and replies from Twitter as dialogs, removing inappropriate dialogs, and constructing a dialog corpus from the remain.

To collect dialogs from Twitter, first we search tweets by a keyword. If the tweet is a reply of another tweet, we retrieve both of them. We repeat this procedure unless the tweet is not a reply of another tweet. In this way, we collect sequence of tweets and replies as a dialog. Finally, we keep the dialog if its length (the number of the tweets) is greater than or equal to 3. The Twitter API is used to search and collect tweets. In order to collect natural dialogs, we use a list of 1,686 words whose word familiarity are high, such as “everyone”, “reunion”, and “lover”, as search keywords.

In the next step, inappropriate dialogs are detected and removed from the collected dialogs. First, we analyze 100 dialogs and investigate what kinds of dialogs are inappropriate, how they are categorized, and how to detect them. As a result, we define four rules to remove inappropriate dialogs. The first one, $\mathbf{R_{short}}$, is a rule that removes dialogs containing a short tweet (utterance). Dialogs including an extremely short utterance are often not real dialogs. We remove dialogs if they contain a tweet that consists of only one Hiragana character (except for interjection), symbols such as punctuation, or emoji. The second one, $\mathbf{R_{line}}$, is a rule that removes dialogs including a tweet with multiple lines. Utterance including multiple lines does not usually appear in a real dialog, although it may appear in sequence of tweets and replies when users make their own stories on Twitter. We remove dialogs if they contain a tweet that fulfills the following conditions: (1) it includes multiple pairs of parentheses (lines are usually indicated by parentheses), (2) the character length in parentheses is more than or equal to 6, (3) a word after the parentheses is not a case marker. The conditions (2) and (3) are set since parentheses are often used not to mark up a line but to emphasize a short noun. The third one, $\mathbf{R_{image}}$, is a rule that removes dialogs including images and URLs. If a tweet contains an image or URL of another web page, people cannot understand a dialog if they do not see and know the contents of the image or the linked web page. We distinguish dialogs where people can not understand them without image or URL and ones where people can understand even without image or URL. We aims at removing only the former. More specifically, we remove a dialog if it contains a tweet including an image or URL and there exists a demonstrative such as "this" or "that" around the image. The presence or absence of a demonstrative is checked to determine whether the tweet mentions the image. The fourth one, $\mathbf{R_{invite}}$, is a rule that removes dialogs if they start with a tweet that widely calls something to other Twitter users. A pseudo dialog is not a real dialog if it includes a call for many people such as Ogiri, which is a game where one user provides a question and other users reply funny answers. A list of users who run Ogiri is manually created in advance. A dialog is removed if the user of the tweet of the beginning of the dialog is included in the list.

Several experiments were conducted to evaluate our proposed method. Dialogs were collected from Twitter between June 8, 2019 and December 25, 2019. The number of collected dialogs was 92,207; the average length was 9.50. Thus we were able to collect a large number of relatively long dialogs. Next, we randomly selected 100 dialogs, and two subjects independently determined whether those dialogs were appropriate or not. The κ coefficient of two subjects was 0.60. In this way, two test datasets were prepared; each is 100 dialogs annotated with the judgment with one subject. Then, the performance of the detection of inappropriate dialog by the proposed method was measured on these two datasets. The precision was 0.75 and 0.75, the recall was 0.32 and 0.43, and the F-measure was 0.45 and 0.55. The major cause of low recall was that many inappropriate dialogs including images were failed to be detected.

Next, we evaluated the individual rules. Note that the rule $\mathbf{R_{invite}}$ was not evaluated because it used the manually created list of Ogiri users. Among dialogs that were judged as inappropriate by each rule, 50 dialogs were chosen as the test data. Two subjects independently judged whether they were inappropriate or not. The κ coefficient of the two subjects were 0.37 for $\mathbf{R_{short}}$, 0.47 for $\mathbf{R_{line}}$, and 0.77 for $\mathbf{R_{image}}$. The precision of $\mathbf{R_{short}}$, $\mathbf{R_{line}}$, and $\mathbf{R_{image}}$were 0.96/0.94, 0.74/0.76, and 0.78/0.78, respectively. It was found that the precision of detection of inappropriate dialogs of the proposed rules was relatively good.

In the future, it is necessary to refine the rules to detect inappropriate dialogs to improve the precision and recall. It is also necessary to investigate another types of inappropriate pseudo dialogs and design methods to automatically remove them.