JAIST Repository
https://dspace.jaist.ac.jp/

Title	小説からの対話コーパスの自動構築
Author(s)	杜，宇龍
Citation	
Issue Date	2019-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/15903
Rights	
Description	Supervisor:白井　清昭，先端科学技術研究科，修士（情報科学）

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Automatic Construction of Dialog Corpus from Novels

1710137 DU,Yulong

In recent years, there are many attempts of research and development of dialog systems. A dialog system is a system that interacts with humans through dialog in natural language. Especially, free conversational systems that can chat with human beings have been paid much attention. However, a large amount of dialog corpora are required for research and development of free conversational systems. A dialog corpus is a database that collects a large amount of dialogs between two or among three or more people. However, it is difficult to construct a large-scale dialog corpus, since it requires much cost to record and transcribe dialogs between humans and to eliminate personal information from dialogs for protection of privacy.

This thesis aims at automatically constructing a large-scale dialog corpus by extracting consecutive utterances made by multiple characters in novels. Utterances in a novel can be regarded as speech by people, and consecutive utterances can be regarded as a dialog. Dialog in a novel is not spontaneous; it is made by an author. However, it is natural as one made by humans. Thus, a dialog corpus excerpt from novels can be useful for developing a free conversational system that realizes natural chat. A wide variety of topics are appeared in novels. Furthermore, the number of novels in the world is quite huge. Therefore, novels are an appropriate information source for automatic construction of a dialog corpus. However, it is insufficient to only extract utterances from a novel. In a dialog corpus, it is required to give a speaker for each utterance. This thesis proposes a method to identify a speaker of utterance in a novel, and extract utterances and their speakers as a dialog corpus. It is the first attempt to automatically construct a dialog corpus from Japanese novels.

In the proposed method, preprocessing is first performed. Metadata other than a text is removed, then a text is split into sentences. Next, utterances are extracted from a novel. By pattern matching with regular expression, sentences in parentheses are extracted as utterances. Next, characters are extracted from a novel. First, words or compound words detected as "person name" by a named entity extraction tool are extracted as characters. CaboCha is used as a named entity extraction tool in this study. In addition, nouns that have a semantic class of people in a thesaurus are also extracted as characters. Specifically, our system extracts nouns in the semantic classes of "person name" and "person" in the Japanese thesaurus Nihongo-goi-taikei.

Next, a speaker is identified for each extracted utterance. We define two types of speakers: an explicit speaker and implicit speaker. An explicit speaker is a speaker who is clearly stated in a novel as he/she says a certain utterance. On the other hand, an implicit speaker is a speaker who is not explicitly indicated in a novel but can be understood by readers that he/she says it. First, patterns to extract explicit speakers are designed. For example, in a pattern "A says that B.", the person A is extracted as a speaker of the utterance B. When an utterance is embedded in a sentence and its speaker is not identified by the above mentioned patterns, it is regarded as not an utterance, although it is once extracted as an utterance. Next, implicit speakers are identified. Several patterns are made to extract characters around an utterance, then they are used to identify an implicit speaker. Finally, speakers are identified by using a speaker alternation pattern when the speaker identification is failed by the pattern matching. The speaker alternation pattern assumes that speakers of consecutive utterances alternate in turn. Using this pattern, speakers are identified by referring a speaker of previous or succeeding utterance that is identified by the patterns to extract explicit and implicit speakers. The speaker alternation pattern is applied repeatedly until speakers of all the utterances are identified, or until speakers are not identified any more. Finally, consecutive utterances with their speakers identified by the above procedures are extracted to construct a dialog corpus. Names of characters in a novel are replaced with speaker IDs such as "speaker A" and are tagged in a dialog corpus, instead of giving a character name as it is.

An experiment was conducted to evaluate our proposed method. Four novels were randomly chosen from novels published in Aozora Bunko as a test data. For each extracted utterance in the test data, its speaker is manually tagged as gold data. The proposed method to identify speakers was evaluated on this test data. The applicability, which was defined as a ratio of the number of utterances that the system can determine the speaker to the total number of speakers, of the proposed method was 1 in all the novels. That is, speakers could be identified for all utterances. On the other hand, the precision for the speaker identification was 0.72. In the novel for which the system most poorly performed, the characters were represented by nicknames, and they could not be extracted as person names by the methods using named entity extraction or thesaurus. If there is a list of characters as metadata of a novel, speakers of utterances can be identified more accurately, resulting improvement of the recall. Finally, dialogs were extracted by the proposed method from 4,836 novels in Aozora Bunko. The applicability of speaker identification was 0.917. The number of dialogs that consists of consecutive utterances where the speakers of all utterances are identified

was about 19,000. The average of the number of utterances per dialog was 10. It indicates that a comparatively large dialog corpus can be constructed automatically by the proposed method.