Contribution ID: **16** Type: **Oral Presentation**

Automatic Extraction of Extended Named Entities from Wikipedia for Conversation Analysis

Thursday, 22 March 2018 12:00 (30 minutes)

In recent years, AI assistants have been developed and these are increasingly helping humans at work and at home. Among them, voice assistances technologies are particularly attractive. They are not limited only to smart phone applications, such as Apple Siri and Google Assistance. Recently Amazon Alexa is used at work and Google Home and LINE Wave are used at home. These voice assistance technologies also help in the business places as well as at home.

As described above, a voice assistant is active in various situations, but there are many operating steps to complete. These steps are quite complicated because they require the use of complex techniques. The steps can be roughly divided into two parts. The first step is to get a voice assistant to understand the human language and to recognize the conversational situation and context. The second step is to create the contents of the next dialogue to continue conversation based on the results obtained in the first step, and then choose a suitable response among from possible choices of responses.

In this research, focusing on the first step we propose a method for capturing and structuring human words in more detail. To explain more concretely, this method automatically extracts the extended named entity from the Wikipedia articles, and it structures the contents of the sentences and the dialogues based on the extracted information.

There are seven types of named entities. They are human, organization, location, date, time, money and rate described by MUC (Message Understanding Conference) and used for classification and analysis of sentences. However, the scope of these classifications is ambiguous, and it does not reach the practical level in actual sentence classification and analysis. Therefore, an extended named entity which improved the named entity has been proposed. Nevertheless, even though it has expanded from seven types to more than 200 types of named entities, it is necessary to deal with vast datasets manually in order to actually perform the classification work. To solve this problem, this research introduces a way to construct datasets of extended named entity from Wikipedia and classify sentences with more detailed granularity.

Primary author: Mr ISHIZUKA, Daiki (Kanazawa Institute of Technology)

Co-author: Prof. NAKAZAWA, Minoru (Kanazawa Institute of Technology)

Presenter: Mr ISHIZUKA, Daiki (Kanazawa Institute of Technology)

Session Classification: Data Management & Big Data Session

Track Classification: Big Data & Data Management