モジュール分割された日本語対話システムの研究

知能情報メディア主専攻　201611350　江畑拓哉

指導教員　Claus Aranha(コンピュータサイエンス専攻)
櫻井鉄也 (コンピュータサイエンス専攻)
北川高嗣 (コンピュータサイエンス専攻)
今倉暁 (コンピュータサイエンス専攻)
二村保徳 (コンピュータサイエンス専攻)
保國惠一 (コンピュータサイエンス専攻))

提出日　2018 年 2 月 1 日

Abstract

In NLP(natural language processing), there is the field, dialogue system. Recently, this system is often built using some systems which solve a small problem for the dialogue system. These systems use various methods such as rule-based, machine learning etc. In this report, I aimed at the dialogue system in Japanese. First, I figured an abstract image for the dialogue system, and next, I studied some required methods for its image mainly using the deep learning. In addition, we will investigate the method of collecting Japanese data, which is indispensable for constructing for the system, and its preprocessing.

In other words, the subjects of this research are the following five themes. 1st, the dealing methods for Japanese data. I observe Japanese utterance data and conversation data from some free data such as the corpus published by NTT(Nippon Telegraph and Telephone Corporation) or Twitter data and consider these natures, some preprocessing methods for them. 2nd, the classification task for imbalanced data. The meaning of “imbalance” indicates that the difference of the abstraction degree of each class and the number o teaching data of each class. 3rd, the conversation model using machine translation model. I tried some machine translation models for conversation model which responses 1-by-1 such as a general greeting.4th, the sentence style translation. I tried several models which I chose to learn translation of sentence style. In Japanese, the change at the sentence end (this research deals with the main agenda of style translation) affects people’s detection of persona too much. So, I believe the style translation will lead this dialogue system user to feel more comfortable by choosing the best style for its system. 5th, the error (in which sentences generated by some deep learning models) detection methods. The main reason of the low of motivation choosing deep learning model to consist a dialogue system is a distrust of precision comparing to rule-based. In this theme, to connect some error handling, I experimented the detection method of unnatural sentence generated by my model which made in the 3rd theme.