# 論文内容の要旨

## 論文題目

Automatic image classification and annotation methods towards development of comprehensive human brain database

(ヒト脳俯瞰データベース構築のための図分類とアノテーション手法の自動化)

氏名　石井 奈都

Due to rapid progress in the development of neuroimaging technologies to analyze human brain activity, many neuroscience research articles are published each year. This phenomenon has created significant demand for a large-scale neuroscience database to store the information in these articles, particularly results regarding the relationship between brain region and function. However, manual curation is extremely labor-intensive and time-consuming, and automatic information extraction program would be very useful.

The aim of this study is to develop a database system that can search and overview neuroscience research outcomes by extracting information from neuroscience research full-text articles. To this end, I decided to index neuroimaging figures in neuroscientific articles with important key terms (e.g. names of brain region, brain function, and task/stimuli given to subjects) from individual articles, because figures in articles usually represent primary research outcomes in a visually understandable way. Such a system would enable researchers to identify various brain regions related to a specific function, along with the tasks and modalities used in relevant studies.

In this thesis, methods for automatic figure classification and annotation are proposed to achieve the above mentioned database system. Regarding figure classification, I developed a text-based figure classification system to automatically select and classify figures of my interest (neuroimaging figures) according to the imaging modality used. This was achieved by Support Vector Machines using word vectors representing each figure in the data set, utilizing texts in figure legends and the main text of the articles. To make the best use of available texts, not only bag-of-words but also bi-grams of words were adopted as features to compose word vectors. This enabled the system to consider multi-word terms of great importance that are frequently used in neuroscience. The

combination of bag-of-words and bi-gram features was quite powerful. As for figure annotation, I developed a key term extraction method to annotate each figure, based on the contents of an article that contains the figure. The methodology proposed in this study was based on graph-based ranking model, termed TextRank, which allows an unsupervised article-level key term extraction. To extract more specific terms to each article, I utilized the statistical information drawn from individual articles as well as other articles in the same and other domains. The method proposed in this thesis outperformed the original TextRank algorithm and other previous methods such as *TF*IDF*. Furthermore, I devised several means to improve the precision of my method. I introduced a new measure to quantify the degree to which a given word sequence is considered to be an actual and meaningful term. I also used the word frequency distributions in the article and articles in other domains to filter out “junk” terms. Finally, among extracted key terms for each article, I identified those representing names of brain regions using statistical information in the corpus so that the users of the database system can specify a region of his/her interest. As a consequence, a prototype database system that can search neuroimaging figure by key terms was implemented, based on these results.

I believe that a system capable of gathering and storing neuroscience research outcomes in an efficient manner will promote research in this area and benefit our understanding of the function of the human brain.