

EXTRACTION AND VISUALIZATION OF NUMERICAL AND NAMED ENTITY INFORMATION FROM A VERY LARGE NUMBER OF DOCUMENTS USING NATURAL LANGUAGE PROCESSING

MASAKI MURATA¹, TAMOTSU SHIRADO¹, KENTARO TORISAWA¹
MASAKAZU IWATATE², KOJI ICHII³, QING MA⁴ AND TOSHIYUKI KANAMARU⁵

¹Language Infrastructure Group, MASTAR Project
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{ murata; shirado; torisawa }@nict.go.jp

²Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
masakazu-i@is.naist.jp

³Graduate School of Engineering
Hiroshima University
1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan
ichiikoji@hiroshima-u.ac.jp

⁴Faculty of Science and Technology
Ryukoku University
Otsu, Shiga 520-2194, Japan
qma@math.ryukoku.ac.jp

⁵Graduate School of Human and Environmental Studies
Kyoto University
Yoshida-Nihonmatsu-cho, Sakyo, Kyoto 606-8501, Japan
kanamaru@hi.h.kyoto-u.ac.jp

Received January 2009; revised May 2009

ABSTRACT. *We have developed a system that semi-automatically extracts numerical and named entity (NE) sets from a large number of Japanese documents by using natural language processing and creates various types of tables and graphs. Our system semi-automatically created approximately 300 types of graphs and tables from newspaper articles collected over a two-year period at accuracy rates of 0.2-0.8, with only two hours of manual work. These newspaper articles contained a large volume of data and not all of them could be read or checked manually in such a short period of time. Consequently, we concluded that our system is useful and convenient for extracting numerical and NE information from a large number of documents. In this paper, we present various types of graphs and tables, covering topics such as accidents caused by cracks in train windows and data on delayed and cancelled trains, that were generated using our system. We also present information about a sample system that extracts text data from web-based news, then derives numerical and NE sets from this text data, and then displays the sets using a graph.*

1. Introduction. Text documents contain many types of numerical and named entity (NE) information such as temperature, humidity and place names. The ability to identify this information and derive graphical representations from it is invaluable in the mining of information from text documents [1,6,24]. In this study, we constructed a system that semi-automatically extracts numerical and NE sets from a large number of Japanese