International Journal of Innovative Computing, Information and Control
Volume **6**, Number **3(B)**, March **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 1549–1568

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

Masaki Murata[1], Tamotsu Shirado[1], Kentaro Torisawa[1]
Masakazu Iwatate[2], Koji Ichii[3], Qing Ma[4] and Toshiyuki Kanamaru[5]

[1]Language 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

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

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

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

[5]Graduate 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

Natural Language Processing (NLP). What it is and why it matters. Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.Â This manual and arduous process was understood by a relatively small number of people. Now you can say, â€œAlexa, I like this song,â€ and a device playing music in your home will lower the volume and reply, â€œOK. Rating saved,â€ in a humanlike voice.Â Royal Bank of Scotland uses text analytics, an NLP technique, to extract important trends from customer feedback in many forms. The company analyzes data from emails, surveys and call center conversations to identify the root cause of customer dissatisfaction and implement improvements. Entity extraction. Natural language processing (NLP), which is the â€œunderstandingâ€ of the natural human language by computers, involves machine translation, information retrieval, and question answering. They are becoming increasingly critical in a variety of applications such as machine reading and understanding, intelligence analysis, social media analysis, etc.Â Most commonly used forms of extractions are entity extraction and their relationship or association extraction.Â Manually tagging entities of all two types in such large set of documents and comparing these entities to those found automatically by the extractor would require a lot of work. Named entity processing has been granted much attention by several speech research groups [1, 2, 3]. For several years now identication of named entities (NE) has been the subject of a lot of academic but also practically oriented research. The importance of the problem led to the need for standardization and motivated extensive work on the named entity denitions.Â We propose a categorization of NE-tasks into detection, lo-calization and value extraction and explain differences and in-terconnections among these subtasks. We also suggest novel methods for solving each of them. In particular, we use the SVM-classier to perform the detection task and, based on its outcome, turn on the localization module implemented as an error-tolerant composition of nite state transducers.