

# Workshop

on

# Natural Language Processing for Asian Languages

October 8, 2009

早稲田大学

長岡技術科学大学

浦野義頼研究室・三上喜貴研究室

本ワークショップに参加ご希望の方は、下記メールアドレスへ事前登録をお願いいたします。

[yekyawthu@aoni.waseda.jp](mailto:yekyawthu@aoni.waseda.jp)

## Welcome to WNLP\_Asian2009

WNLP\_Asian2009 focuses on Natural Language Processing (NLP) of Asian languages. This workshop is a part of GITS/GITI Research Festival 2009, and jointly launched between Prof. Yoshiyori Urano Lab., Graduate School of Global Information and Telecommunication Studies (GITS), Waseda University and Prof. Yoshiki Mikami Lab., Nagaoka University of Technology. NLP research in progress will be presented for Myanmar, Uyghur, Thai and Chinese languages followed by discussions. This workshop, open to anyone interested without registration fee, will be held at Honjo campus, GITS, Waseda University, Honjo City.

Let's move towards NLP development of Asian languages!

Ye Kyaw Thu

WNLP\_Asian2009 Organizer

yekyawthu@aoni.waseda.jp



Global Information and Telecommunication Studies (GITS), Waseda University  
Address: 1011 Nishi-Tomida, Honjo-shi, Saitama 367-0035 JAPAN  
Phone: 0495-24-6420



Prof. Yoshiyori URANO, *GITS, Waseda University*

Yoshiyori Urano received B.E., M.E. and Doctor of Engineering from Waseda University, in 1960, 1962 and 1965 respectively. Dr. Urano joined KDD (now KDDI) in 1965, served as Director of KDD Research and Development Laboratories in 1993-1996, and moved to Waseda University in 1996. He is now a Professor, Graduate School of Global Information and Telecommunication Studies, Waseda University. His current interests include Next-Generation Internet, Information Network Architecture, Intelligent Network Operation/Management, Multimedia, Distributed Processing, Network Applications (Distance Education/e-Learning/Ubiquitous Learning, Tele-medicine Information System, Disaster Prevention Information System, etc.)



Prof. Yoshiki MIKAMI, *Nagaoka University of Technology*

Yoshiki Mikami is a professor of management and information science at Nagaoka University of Technology, Japan. He has initiated several international collaborative projects, such as Language Observatory Project, the Asian Language Resource Network Project, and the Country Domain Governance Project. He also serves as a chairman of Joint Advisory Committee for ISO registry of character codes. Mikami received a BE in mathematical engineering from the University of Tokyo and a PhD from the Graduate School of Media and Governance at Keio University.

10:00-10:30

### Opening Plenary

Prof. Yoshiyori URANO, *GITS, Waseda University*

10:30-11:00

### A Proposal for Formal Description Method of Collating Orders

Yoshiki Mikami, Wunna Ko Ko, *Nagaoka University of Technology*

Sort orders vary culture to culture. Authentic dictionaries and national level language committees have defined standard sort orders, or collating sequences for their languages. The rules set in these dictionaries or by committees are, however, complex and are quite difficult for foreigners to understand them precisely. In response to these difficulties, generalized description of sort orders have been developed along with the spreading usage of multi-lingual applications. Unicode Standard (1991), Canadian standard CSA Z243.4.1 (1992), Japanese standard JIS X 4061 (1996) are a few of those efforts. The paper proposes a formal description method of collating orders by employing a concept of partially ordered set (POSET). Also the usefulness of this method is demonstrated by applying it to Myanmar language collating order.

11:00-11:30

### Measuring Phonetic Similarities of Words Across Languages

Ohnmar Htun, Shigeaki Kodama, Yoshiki Mikami, *Nagaoka University of Technology*

The objective of this study is to propose a methodology to measure the phonetic similarities between words across languages. The basic idea behind the study is Soundex algorithm. We modified the original Soundex algorithm to make it applicable for cross language comparison and introduced Levenstein distance to give a quantitative measure of phonetic similarities. In the cyberspace, we find a lot of examples of phonetically translated words; technical terms, names of places and persons, etc. If we find a good measure of phonetic similarities of words expressed in different languages, the measure can be used for various applications such as search engines, online dictionaries, etc. The usefulness of the proposed measurement methodology is demonstrated by applying it to the analysis of multilingual terminology dictionary data, which covers more than 10,000 science and engineering terms translated into eleven Asian languages (which are also developed by authors). Our study shows that this methodology can help us to identify cognates across different languages automatically.

**11:30-12:00**

### **Language Specific Crawler for Myanmar Web Pages**

Pann Yu Mon, Chew Yew Choong, Yoshiki Mikami, *Nagaoka University of Technology*

With the enormous growth of the World Wide Web, Search engines play a critical role in retrieving information from the borderless Web. Although many search engines are available for the major languages, they are not much proficient for the less computerized languages including Myanmar because of the lack of standard encodings. Due to the use of multiple non-standard encodings, most of the search engines cannot detect the Myanmar Web pages correctly. In this paper, a language specific crawler (LSC) which is implemented as multi-threaded objects that run concurrently with language identifier is presented. It is capable of collecting the Myanmar Web pages as many as possible for the multiple encodings. It is also proved that the implemented algorithm can be able to collect the Myanmar pages at a satisfactory level of coverage. Here the evaluation of crawler by means of two standard criteria such as Recall and Precision is discussed. And it is also discussed how to measure the total numbers of the Myanmar Web pages on the entire Web as extra experiment. In the end the experiments that were conducted along with the coverage of LSC on Myanmar language Web content is accordingly reported.

**13:30-14:00**

### **Analysis on Possible Combinations of a Consonant with Vowels and Tone Marks for Thai**

Jakchai Butsrkui, Ye Kyaw Thu, Mitsuji MATSUMOTO, Yoshiyori URANO, *GITS, Waseda University*

In order to design the Positional Prediction text input method for Thai language, we analyze the possible combinations of a consonant with vowels and tone marks. This analysis is based on grammatical syllable formations and pronounceable combinations. We consider all of the combinations to form meaningful word or a part of it. We present the possible combination patterns as well as the number of possible combinations of each consonant with vowels and tone marks of current Thai language.

**14:00-14:30**

### **Myanmar Web Pages Indexer with an Enhanced Dictionary Method**

Pann Yu Mon, Chew Yew Choong, Yoshiki Mikami, *Nagaoka University of Technology*

Search Engine, the most effective Information Retrieval tool (IR) plays a critical role in helping user to find correct information from the borderless Web. Although many Search Engines are available for the popular World languages (for example English, French, etc.), but they are not much proficient for the majority of vernacular languages in the world. The Myanmar language, likes many other less popular languages, is not within the language preference of all major Search Engines. Therefore, a Search Engine that is capable of searching the Web documents written in any language rather than in popular languages is highly needed, especially when more and more Web sites are coming up with localized content.

In this study, the design and the architecture of a dictionary based indexer for Myanmar Web pages is proposed and which are expected to be used in Search Engines for Myanmar language.

**14:30-15:00**

### **Prototyping with "Pictures & Mobile Devices": A Rapid Prototyping Technique for Mobile User Interfaces**

Ye Kyaw Thu, Yoshiyori URANO, *GITS, Waseda University*

Low-fidelity prototyping is an important tool for developers and designers to test their user interfaces and solutions at the early stage of the product life cycle. In these days, the method has also been adapted and widely used for mobile interface and interaction development. This paper introduces a rapid prototyping technique with pictures and mobile devices. This prototyping process is based on the traditional paper prototyping technique. The main difference is that we draw screen design or user interface for mobile devices with drawing software and display it on the screen of real mobile devices for user evaluation. We discuss the creation process and use of "Pictures & Mobile Devices" prototyping for developing two Myanmar text input interfaces. Our approach can reveal many usability problems that should be found at the early stage of mobile user interface design.

**15:00-15:30**

### **Coffee Break**

**15:30-16:00**

### **Romanized Myanmar Language Input System for Mobile Phone**

Hlaing Myint Oo, Katsuko T. Nakahira, Yoshiki Mikami, *Nagaoka University of Technology*

Romanized Myanmar language input method for mobile phone was introduced in order to make user no need to memorize the character positions on keypad leading widespread use of Myanmar-based SMS. As there is no efficient database for input system, a database is created by making survey on native speakers and analyzing the result in probabilistic method. In this paper, we discuss about the procedures in developing database. We develop the prototype for the input system and make evaluation and comparison among other existing input systems.

**16:00-16:30**

### **Uyghur Character Code and Text Processing**

Omerjan Osman, Yoshiki Mikami, *Nagaoka University of Technology*

Characters used by Uyghur people have changed over time, from Orkhon script in the ancient times to currently used Arabic script. Since 8th century until 18th century Sogd-origin script, we call it just “Uyghur script”, had been used. Authors have developed a proposal to ISO/IEC 10646 based character code and a text processing system for the Uyghur script based on a manuscript written in the Uyghur script. While manuscripts and documents written in the Uyghur script can be found in museums and universities around the world, most of them are just kept in their original form and only a few of them are electronically scanned as an image file and archived and have never been handled as coded text data. Therefore, no quantitative analysis has been done, in spite of their rich cultural value. In this research, based on the Uyghur character code table which we are suggesting to Unicode and ISO/IEC10646, we have designed Uyghur TrueType Font and developed a text input system based on a generic I/O system developed by Tokyo University of Foreign Studies.

**16:30-17:00**

## **Archiving Web Contents of Significant Events in Myanmar**

Zaw Zaw Aung, *Nagaoka University of Technology*

The Internet Archive (<http://www.archive.org>) is working to prevent the Internet - a new medium with major historical significance - and other "born-digital" materials from disappearing into the past. It provides "Archive-It" service that allows institutions to build and preserve their own web archive of born digital content, through a user friendly web application, without requiring any technical expertise or hosting facilities. Subscribers can harvest, catalog, and archive their collections, and then search and browse the collections when complete. Collections are hosted at the Internet Archive data center, and accessible to the public with full text search. Additionally, in collaboration with some institutions, they do the special web collections for world events, disasters, etc. Examples include Asian Tsunami Web Archive, Hurricanes Katrina and Rita, Election 2002, September 11th and many others. These collections are browsable by category and url, and/or searchable by customized search engine NutchWAX. However, language support for archiving and searching is still lacking and their primary focus is on U.S. related events. Each and every country has its own significant event in history. Examples include Myanmar 2007 Suffron Uprising, 2008 Nargis Cyclone, 2008 China Earthquake. For such events, people emotion, real rescue effort, the stress condition and the volunteer's impressions are closely expressed in personal web blogs, local news agencies and online forums which are written in own national language. Therefore, this discussion section will provide facts and feasibility of developing language/event specific archiving, categorizing and searching with some potential approaches.

**17:00-17:30**

## **A Proposal for a Myanmar Language Stemmer**

San Ko Oo and Yoshiki Mikami, *Nagaoka University of Technology*

Stemming is a process of removing affixes (prefixes and suffixes) in a word to generate a root word. In this paper, we proposed a stemmer used for Myanmar language. That stemmer is applied for an individual word by stripping a suffix and a prefix to produce the correct root word based on the grammatical categories. It is expected that if this procedure is fully developed, the stemmer will result in Myanmar Information Retrieval.

**17:30-18:00**

**The Development and Evaluation of Chinese Four Tones Discrimination CAI System**

Song Liu, Yoshiyori URANO, *GITS, Waseda University*

A CAI system for self-teaching of discriminating Chinese four tones has been developed and provided through the Internet to Japanese college students of Chinese language class. In this report, efficiency of this system is discussed, based on the analysis of the characteristic errors by about 100 students from the results of two experimental examinations.

**18:00-18:30**

**Closing Plenary**

Prof. Yoshiki MIKAMI, *Nagaoka University of Technology*