IJCNLP-08 Workshop

On

NER for South and South East Asian Languages

Proceedings of the Workshop

12 January 2008
IIIT, Hyderabad, India

Introduction

Welcome to the IJCNLP Workshop on Named Entity Recognition for South and South East Asian Languages, a meeting held in conjunction with the Third International Joint Conference on Natural Language Processing at Hyderabad, India. The goal of this workshop is to ascertain the state of the art in Named Entity Recognition (NER) specifically for South and South East Asian (SSEA) languages. This workshop continues the work started in the NLPAI Machine Learning Contest 2007 which was focused on NER for South Asian languages. NER was selected this time for the contest as well as for this workshop because it is one of the fundamental and most important problems in NLP for which systems with good accuracy have not been built so far for SSEA languages. The primary reason for this is that the characteristics of SSEA languages relevant for NER are different in many respects from English, on which a lot of work has been done with a significant amount of success in the last few years. An introductory article further explains the background of and motivation for this workshop. It also presents the results of an experiment on a reasonable baseline and compares the results obtained by the participating teams with the results for this baseline.

The workshop had two tracks: One track for regular research papers on NER for SSEA languages and the second track on the lines of a shared task. The workshop attracted a lot of interest, especially from the South Asian region. Participation from most of the research centers in South Asia working on NER ensured that the workshop met its goal of ascertaining and advancing the state of the art in NER for SSEA languages. Another major achievement was that a good quantity of named entity annotated corpus was created in five South Asian languages. The notable point about this effort was that this was done almost informally on a voluntary basis, without funding. This is an important point in the context of SSEA languages because lack of annotated corpora has held back progress in many areas of NLP so far in this region.

Each paper was reviewed by three reviewers to ensure satisfactory quality of the selected papers. Another major feature of the workshop is that it includes two invited talks by senior researchers working on the NER problem for South Asian languages. The only drawback of the workshop was that there was no paper on any South East Asian language.

We would like to thank the program committee members for all the hard work that they did during the reviewing process. We would also like to thank all the people involved in organizing the IJCNLP conference. We hope that this workshop will help in creating interest in NER for SSEA languages and we will soon be able to achieve results comparable to those for languages like English.

Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh (Chairs)

Organizers:

Rajeev Sangal, IIIT, Hyderabad, India
Dipti Misra Sharma, IIIT, Hyderabad, Hyderabad, India
Anil Kumar Singh, IIIT, Hyderabad, India

Program Committee:

Rajeev Sangal, IIIT, Hyderabad, India
Dekai Wu, The Hong Kong University of Science & Technology, Hong Kong
Ted Pedersen, University of Minnesota, USA
Dipti Misra Sharma, IIIT, Hyderabad, Hyderabad, India
Virach Sornlertlamvanich, TCL, NICT, Thailand
Alexander Gelbukh, Center for Computing Research, National Polytechnic Institute, Mexico
M. Sasikumar, CDAC, Mumbai, India
Sudeshna Sarkar, Indian Institute of Technology, Kharagpur, India
Thierry Poibeau, CNRS, France
Sobha L., AU-KBC, Chennai, India
Tzong-Han Tsai, National Taiwan University, Taiwan
Prasad Pingali, IIIT, Hyderabad, India
Canasai Kreungkrai, NICT, Japan
Manabu Sassano, Yahoo Japan Corporation, Japan
Kavi Narayana Murthy, University of Hyderabad, India
Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India
Anil Kumar Singh, IIIT, Hyderabad, Hyderabad, India
Doaa Samy, Universidad Autnoma de Madrid, Spain
Ratna Sanyal, IIIT, Allahabad, India
V. Sriram, IIIT, Hyderabad, Hyderabad, India
Anagha Kulkarni, Carnegie Mellon University, USA
Soma Paul, IIIT, Hyderabad, Hyderabad, India
Sofia Galicia-Haro, National Autonomous University, Mexico
Grigori Sidorov, National Polytechnic Institute, Mexico

Special Acknowledgment:

Samar Husain, IIIT, Hyderabad, India
Harshit Surana, IIIT, Hyderabad, India

Invited Speakers:

Sobha L., AU-KBC, Chennai, India
Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India

Table of Contents

Invited Talk: *Named Entity Recognition: Different Approaches*
Sobha L 1

Invited Talk: *Multilingual Named Entity Recognition*
Sivaji Bandyopadhyay 3

Named Entity Recognition for South and South East Asian Languages: Taking Stock
Anil Kumar Singh 5

A Hybrid Named Entity Recognition System for South and South East Asian Languages
Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar and Pabitra Mitra . 17

Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition
Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma 25

Language Independent Named Entity Recognition in Indian Languages
Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay . . 33

Named Entity Recognition for Telugu
P Srikanth and Kavi Narayana Murthy 41

Bengali Named Entity Recognition Using Support Vector Machine
Asif Ekbal and Sivaji Bandyopadhyay 51

Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields
Vijayakrishna R and Sobha L 59

A Character n-gram Based Approach for Improved Recall in Indian Language NER
Praneeth M Shishtla, Prasad Pingali and Vasudeva Varma 67

An Experiment on Automatic Detection of Named Entities in Bangla
Bidyut Baran Chaudhuri and Suvankar Bhattacharya 75

Hybrid Named Entity Recognition System for South and South East Asian Languages
Praveen P and Ravi Kiran V 83

Named Entity Recognition for South Asian Languages
Amit Goyal 89

Named Entity Recognition for Indian Languages
Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanyal 97

Experiments in Telugu NER: A Conditional Random Field Approach
Praneeth M Shishtla, Karthik Gali, Prasad Pingali and Vasudeva Varma 105

Workshop Program

Saturday, January 12, 2008

Session 1:

09:00-9:30 **Opening Remarks:** *Named Entity Recognition for South and South East Asian Languages: Taking Stock*
Anil Kumar Singh

09:30-10:00 **Invited Talk**: *Named Entity Recognition: Different Approaches*
Sobha L

10:00-10:30 *A Hybrid Named Entity Recognition System for South and South East Asian Languages*
Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar and Pabitra Mitra

10:30-11:00 **Break**

Session 2:

11:00-11:30 **Invited Talk**: *Multilingual Named Entity Recognition*
Sivaji Bandyopadhyay

11:30-12:00 *Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition*
Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma

12:00-12:30 *Language Independent Named Entity Recognition in Indian Languages*
Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay

12:30-14:00 **Lunch**

Session 3:

14:00-14:30 *Named Entity Recognition for Telugu*
P Srikanth and Kavi Narayana Murthy

14:30-15:00 **Poster Display and Discussion**

An Experiment on Automatic Detection of Named Entities in Bangla
Bidyut Baran Chaudhuri and Suvankar Bhattacharya

Saturday, January 12, 2008 (continued)

Hybrid Named Entity Recognition System for South and South East Asian Languages
Praveen P and Ravi Kiran V

Named Entity Recognition for South Asian Languages
Amit Goyal

Named Entity Recognition for Indian Languages
Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanyal

Experiments in Telugu NER: A Conditional Random Field Approach
Praneeth M Shishtla, Karthik Gali, Prasad Pingali and Vasudeva Varma

15:30-16:00 **Break**

Session 4:

16:00-16:30 *Bengali Named Entity Recognition Using Support Vector Machine*
Asif Ekbal and Sivaji Bandyopadhyay

16:30-17:00 *Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields*
Vijayakrishna R and Sobha L

17:00-17:30 *A Character n-gram Based Approach for Improved Recall in Indian Language NER*
Praneeth M Shishtla, Prasad Pingali and Vasudeva Varma

17:30-18:00 **Closing Discussion**