Abstract

Machine transliteration has received significant research attention in recent years. In most cases, the source language has been English and the target language is an Asian language. This paper focuses on Hindi to English machine transliteration of Indian named entities such as proper nouns, place names and organization names using conditional random fields (CRF). Hindi is the national language of India and spoken by more than 500 million Indians. Hindi is the world’s fourth most commonly used language after Chinese, English and Spanish. This system takes Indian place name as an input in Hindi language using Devanagari script and transliterates it into English. The input to the system is provided in the form of syllabification in order to apply the n-gram techniques. As more than 50% named entities are formed as a combination of two and three syllabic units, the n-gram approach with unigrams, bigrams and trigrams of Hindi are used to train the corpus. The system provides the satisfactory performance for trigrams as compared to unigrams and bigrams.

References
Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields

- Stalls Bonnie Glover and Kevin Knight, Translating names and technical terms in Arabic text. 1998
- Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, and Pabitra Mitra, Named entity recognition in Hindi using maximum entropy and transliteration, 2008

- A Ekbal and S. Bandyopadhyay, Bengali named entity recognition using support vector machine, in Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, pp. 51–58, January 2008
- A Ekbal and S. Bandyopadhyay, Bengali Named Entity Recognition using Classifier Combination, in Proceedings of Seventh International Conference on Advances in Pattern Recognition, pp. 259–262, 2009
- A Ekbal and S. Bandyopadhyay, Voted NER system using appropriate unlabelled data, in Proceedings of the Named Entities Workshop, ACL-IJCNLP 2009,
- Charles Sutton and Andrew McCallum, An Introduction to conditional random fields for relational learning, University of Massachusetts, USA

- http://www. whereincity. com/babynames
- http://www. indianchild. com/

Index Terms

Computer Science  Image Processing

Keywords

Bigram  Conditional Random Fields  Trigram  Transliteration  Syllabification