11-712: NLP Lab Report

Rajarshi Das

Friday 17th January, 2014

Abstract

This is a report on the development of an open source dependency parser for the language, Bengali. Presently I have reported some basic information about the language.

The goal of this project is to design, implement and evaluate a dependency parser for the language, Bengali (also my native language). This language is characterized by a rich system of inflections, derivation and compound formation (Chakroborty, 2003; Saha et al., 2004) which makes analysis and generation of Bengali, a challenging task (Ghosh et al., 2009).

1 Basic Information about Bengali

According to (Lewis, 2013), Bengali is an eastern Indo-Aryan Language and is native to the region of eastern south Asia. It is the official language of Bangladesh and is also spoken in the Indian state of West Bengal and parts of Tripura and Assam.

Bengali follows the SOV order in terms of ordering of subject, object and verb (Dasgupta, 2003). It makes use of postpositions instead of prepositions. Determiners follow the noun while numerals, adjectives and possessors precede the noun. It exhibits no case or number agreement and no grammatical gender phenomena (Dasgupta, 2003). Nouns and pronouns are declined into four cases - nominative, objective, genitive and locative (Bhattacharya, 2001)

Bengali is written using the Bengali script. It has 11 vowel graphemes and 39 graphemes representing consonants and other modifiers. The script is written and read horizontally from left to right. Figure 1 and 2 show the vowels (and its various diacritics) and consonants in the Bengali script (Image source: Internet).

Figure 1: Vowels and vowel diacritics in Bengali script.

Figure 2: Consonants in Bengali script.

References

Tanmoy Bhattacharya. Bengali. In Jane Garry and Carl Rubino, editors, *Facts About the World's Languages: An Encyclopedia of the World's Major Languages: Past and Present*. H.W. Wilson Press, New York/Dublin, 2001.

Bamondeb Chakroborty. Uchchotoro bangla byakaron. *Akshay Malancha*, 2003.

Probal Dasgupta. Bangla. In George Cardona and Dhanesh Jain, editors, *The Indo-Aryan Languages*, pages 351–390. Routledge, London, 2003.

Aniruddha Ghosh, A Das, P Bhaskar, and Sivaji Bandyopadhyay. Dependency parser for bengali: the ju system at icon 2009. *NLP Tool Contest ICON*, 2009.

M. Paul Lewis. Ethnologue: Languages of the world, 2013. URL `http://www.ethnologue.com/`.

Goutam Kumar Saha, Amiya Baran Saha, and Sudipto Debnath. Computer assisted bangla words pos tagging. In *Proceedings of the International Symposium on Machine Translation NLP and TSS (iTRANS)*, 2004.