README

IL-POST Data Version 1.0

Microsoft Research India Pvt. Ltd.
2010

GOAL

To support the task of Part-of-Speech Tagging (POS) and other forms of data driven linguistic research on Indian Languages in general, MSR India has developed POS labeled data for Hindi, Bangla and Sanskrit as a part of the Indian Language – Part-of-Speech Tagset (IL-POST) project.

CORPUS DESCRIPTION

This corpus is designed for those who need annotated text corpora for their work. The corpus is designed based on the IL-POST framework. IL-POST is a POS-tagset framework for Indian Languages, which has been designed to cover the morph-syntactic details of Indian Languages. It supports a three-level hierarchy of Categories, Types and Attributes .The corpus mainly consist therefore of two different level of information for each lexical token (a) lexical Category and Types , and (b) set morphological attributes and their associated values in the context.

Example:

> মুখিশপুরে\NP.0.loc.n.n লোহগড়\NP.0.0.n.n নামে\PP.0.n একটি\JQ.y.n.crd দুর্গ***NC.sg.0.n.n*** গড়ে***VM.0.0.0.0.nfn.n.n.n*** তোলেন\VAUX.3.pst.sim.dcl.fin.n.n.n ।***PU***

The tag follows the word separated by a '\' (back slash) immediately after the word. There are no blank spaces in between. After the whole POS tag there should be at least one blank (white space) before the next word or a sentinel. In the above example, the first string of 2 to 4 uppercase characters denotes the Category and Type. For example, in the above sentence the word দুর্গ *is marked* as **NC** which *stands for Noun Common (***N** *denotes Category Noun and* **C** *denotes type Common).*

The attributes are denoted as numbers or letters, as the case may be, after the tag for the lexical category separated by '.' (dot). The order of the attributes is fixed and cannot be arbitrarily swapped. To illustrate this, consider the category *proper noun* (NC) whose attribute set is {Number, Case-marker, Definiteness, and Emphatic}. *Number* can take values from the set {Singular (sg), Plural (pl), Not-applicable (0)}; *Case-marker* can take values from the set {Accusative (acc), Genitive (gen), Locative (loc), Not-applicable (0)}; *Definiteness* can take values from {yes(y) and no(n)} and *Emphatic* can take values from { yes(y) and no(n)}. *Therefore, for the Common Noun দুর্গ, in the above example sentence, which is singular, not-applicable, non-definite and non-emphatic, the comple tag should be:*

\\NC.sg.0.n.n

- **Corpus size**
    - **Bangla – Manually annotated 7168 sentences (102933 words)**
- Format of the Data

  - The annotated data is available in two folders **Bangla1 (3684 sentences, 51091 words)** and **Bangla 2 (3484 sentences, 51842 words).** The two folders separate the two stages in which the data was annotated.

  - All annotated data is in both XML and TEXT format for all the languages

  - Each data file contains between 3- 5,000 words and kept in sentence level.

  - The XML file contains the metadata about the language, encoding, data size etc.

- This data was created under the supervision of **Multilingual Systems Group, Microsoft Research Labs India**

SOURCE DATA

The Bangla annotated data targets to cover written modern standard Bangla from various sources. **Bangla data is mainly taken from the following resources:**

- Blogs
- Multikulti (**http://www.multikulti.org.uk)**
- Wikipedia
- A portion of CIIL corpus

ANNOTATION PROCEDURE

The detail of the annotation procedure is downloadable along with the annotated data. Please go through the *annotation guideline* (specific to the language) for clarification and further annotation.

DIRECTORY STRUCTURE

In the *MSRI-Data\Annotated_data\Bangla*:

\Bangla1 and \Bangla2 contain:

*.xml files are in the *Annotated_data\XML_files*

*text files are in *Annotated_data\text_files*

In the docs\ directory:

More detailed information about the part-of-speech tagset and annotation process.

CONTACT:

ilpost@microsoft.com