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:

मित्र\NC.neu.sg.voc.viii !\PU पश्य\V.ppd.sg.2.imp.n मे\PPR.0.sg.1.gen.vi.n.n.prx
बुद्धिबलम्\NC.neu.sg.acc.ii ।\PU

सः\PPR.mas.sg.3.nom.i.n.n.dst अब्रवीत्\V.ppd.sg.3.imprf.n -\PU “\PU स्वामिन्\NC.mas.sg.voc.viii !\PU
न\CNG एतेषाम्\PPR.mas.pl.3.gen.vi.n.n.prx एषः\DAB.mas.sg.3.0.i.prx.n दोषः\NC.mas.sg.0.i ।\PU

हिरण्यकः\NP.mas.sg.nom.i अपि\CAD सहस्रमुखबिलदुर्गम्\NC.mas.sg.acc.ii प्रविष्टः\KDP.mas.sg.0.i सन्\CPP
अकुतोभयः\JJ.mas.sg.nom.i.n.n सुखेन\CAD आस्ते\V.ppd.sg.3.prs.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 or 3 uppercase characters denotes the Category and Type. For example, in the above sentence the word बुद्धिबलम्is *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 *common noun* (NC) whose attribute set is {Gender, Number, Case, and vibhakti (nominal declension). *Gender* can take values from {Masculine(mas), Feminine(fem), Neuter(neu), Not-applicable(0)}; *Number* can take values from the set {Singular (sg), Dual (du), Plural (pl), Not-applicable (0)};*Case* can take values from set {Nominative (nom), Accusative (acc), Instrumental (ins), Dative (dat), Ablative (abl), Genetive (gen), Locative (loc), Vocative (voc), Not-applicable(0)}; and *Vibhakti* can take values from the set {Prathama (i), Dwitiya (ii), Tritiya (iii), Chaturthi (iv), Panchami (v), Shashthi (vi), Saptami (vii), Vocative (viii) , Not-applicable (0)}.*Therefore, for the Common Noun बुद्धिबलम्, in the above example sentence, which is neuter, singular, accusative and ii, the comple tag should be:*

\NC.neu.sg.acc.ii

- **Corpus size**
  - **Sanskrit – Manually annotated 3703 sentences (57218 words)**
- Format of the Data
  - The annotated data is available in both XML and TEXT format for both the language
  - There are two data files containing approximately 12,000 and 45,000 words respectively 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
  - **Dr. Girish Nath Jha, Jawaharlal Nehru University, New Delhi.**

SOURCE DATA

The Sanskrit annotated data are collection of *Panchatantra* stories. The first 3 *tantras* and the *Prastavana* part have been covered.

ANNOTATION PROCEDURE

The detail of the annotation procedure is described in the accompanying paper:

Girish Nath Jha, Madhav Gopal, and Diwakar Mishra, "**Annotating Sanskrit corpus: adapting IL-POSTS**" paper presented at the 4th Language Technology Conference, Poznan, Poland. Nov 2009.

DIRECTORY STRUCTURE

In *MSRI-Data\Annotated_data\Sanskrit*:

*.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