9.5 Language Technology Resources

Language Technology Resources -An Overview

1. Preamble

Technology Development in Indian Languages initiative of Department of Information Technology, Ministry of ICT or Government of India has been facilitating and supporting the development of language technology resources in all Indian languages, and promoting their dissemination. Towards this goal it established several Resource Centers for Indian Language Technologies in different Universities and Institutes. A few other funding agencies are also supporting the language technology efforts. Several individuals in India and abroad, and some industries have also been developing different language products/resources. As there are so many distributed efforts and with so many languages it is very difficult to keep track of the activities in this area. It is felt necessary to evolve an indexing system for the language technology resources, so that a suitable portal can be created which can offer a variety of services to persons who are concerned with Indian language technologies. This document presents such an indexing system to keep track of Language Technology Resources. If all the resource centres can follow this indexing and share their documents with TDIL, it would be easy to keep track of the progress and would be possible to enthuse more people to participate in the development activities.

2. Language Technology Resources

The indexing methodology followed is based on the system presented by Hamish Cunningham (2000). Language Technology Resources, in the Indian context, are presented in hierarchical manner in the following. The diagrammatic version of the Language Technology Resources is given in the annexure as figure 1.

1. Enabling Resource

1.1 Input method

1.1.1 Keyboard Driver

1.2 Output method

1.2.1 Fonts

1.2.2 Rendering Engine

1.3 Representation (UNICODE, ISCII, Proprietary)

2. Language Resource

2.1 Lexical resource .r
2.2.1 Term Bank

2.2.2 Thesaurus

2.2.3 Dictionary

2.2 Linguistic Analyzer

2.2.1 Phonetic Analysis

2.2.2 Phonological Analysis

2.2.3 Morphological Analysis

2.2.4 Syntactic Analysis

2.2.5 Semantic Analysis

2.2.6 Discourse Analysis

2.3 Fommlism

2.4 Corpora

2.4.1 Text Corpora

2.4.2 Speech Corpora

2.5 Ontology

3 Processing Resource

3.1 Retriever

3.1.1 Search Engine

3.2 Recognizer

3.2.1 Handwriting Recognizer

3.2.2 Speech Recognizer

3.2.3 Optical Character Recognizer

3.2.4 Speaker Determiner

3.3 Translator

3.4 Generator

3.4.1 Content determiner

3.4.2 Discourse Planner

3.4.3 Speech Synthesizer

3.4.4 Surface Realizer

3.4.5 Lexicalizer

3.5 Analyzer

3.5.1 Super Tagger

3.5.2 POS Tagger

3.5.3 Sense Tagger

3.5.4 Morphological Parser

3.5.5 Syntactic Parser

3.5.6 Semantic Parser

3.5.7 Word Segmenter

3.5.8 Sentence Segmenter

3.5.9 Word Predictor

3.5.10 Phrase Predictor

3.5.11 Text Structure Analyzer

3.5.12 Discourse Analyzer

3.5.13 Lexical Analyzer

3.5.14 Grammar Checker

3.5.15 Summarizer

3.5.16 Spell Checker

3.5.17 Semantic Analyzer

3.5.18 Morpher

3.5.19 Stemmer

3.5.20 Tokenizer

3.5.21 Indexer

3. Document Indexing

The documents on resources generated can be indexed as follows.

The header will consist of eight fields:

Language: A two-letter code is used to represent the language

AS = Assamese

BE = Bengali

EN = English

GU = Gujarati

HI = Hindi

KA = Kannada

KO = Konkani

KS = Kashmiri

MA = Malayalam

MR = Marathi

NE = Nepalese

OR = Oriya

PU = Punjabi

SA = Sanskrit

SI = Sindhi

TA = Tamil

TE = Telugu

UR =Urdu

Resources can be organized hierarchically at multiple levels. As technology progresses more resources may be identified or more levels may be created. The indexing system suggested below would enable us conveniently add items as per the need.

Level O Resource: There are three types of resources

ER: Enabling resources

LR: Language resources (Refers to data-only resources such as lexicons, corpora, thesauri or ontologies. LRs can be on-line or off-line resources.)

PR: Processing Resources (Refers to resources whose character is principally programmatic or algorithmic, such as lemmatisers, generators, translators, parsers or speech recognizers. For example, a part-of-speech tagger is best characterized by reference to the process it performs on text. PRs typically include LRs, e.g. a tagger (often has a lexicon: a word sense disambiguator uses a dictionary or thesaurus.)

Level 1 Resources: There are several sub-types in each type of resource

INP: Input Method (An input method is a program that allows computer users to enter complex characters and symbols by using a standard keyboard.)

OUT: Output Method (Open Type Font is a cross platform font file format developed jointly by Adobe Systems Incorporated and Microsoft. Based on

Unicode standard it is an extension of the TrueType SFNT format that can now support Post Script font data and new Typographic features.)

REP: Representation

LEX: Lexical resource (Resources like dictionary, thesaurus, term bank etc. which contain information regarding a lexicon entry for each word. It contain typical information like part of Speech, inflection class, etc.)

LIA: Linguistic Analysis (Analysis of language at various levels, like phonological level, semantic level etc.)

FOR: Formalism (Refers to the theoretical framework used for linguistic analysis)

CRP: Corpora (It is a collection of writings or recorded remarks used for linguistic analysis e.g. British National Corpus (BNC))

ONT: Ontology (An explicit formal specification of how to represent the objects,concepts and other entities that are assumed to exist in some area of interest and the relationship that hold among them. It is the hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities.)

RTR: Retriever (A tool which retrieves the required item from a given database)

REC: Recognizer (Recognizes an item from the description/representation it is given)

TRA: Translator (Translates the input from one language to another)

GEN: Generator (A program that produces specific programs from the definition of an operation.)

ANL: Analyzer (Analysis of language at various levels, like phonological level, syntactic level, semantic level etc.)

Level 2 Resources: There are several sub-sub-type resources under sub.-type resources

Input Method

KBD: Keyboard Driver

Output Method

FNT:Fonts

REN: Rendering Engine

Representation

UNI: Unicode

ISC: ISCII (Indian Script Code for Information Interchange (ISCI1). It is an encoding format evolved by Department of Electronics intended for use in all computer & communication media which allow usage of 7 or 8 bit characters).

PRP: Proprietary

Lexical Resources

TBN: Term Bank (A stock of terms used in a particular profession, subject, or style: a vocabulary: the lexicon of surrealist art.)

THE: Thesaurus (A Thesaurus is a book of selected words or concepts, such a specialized vocabulary items of a particular field, as of medicine or music. It often contains synonyms, and other semantically related words including related and contrasting words and antonyms.)

DCT: Dictionary (A reference book containing an alphabetical list of words, with information given for each word, usually including meaning, pronunciation and etymology. It lists the words of a language with translations into another language. Electronic dictionaries contain a list of words stored in machine-readable form for reference, as by spelling-checking software)

Linguistic Analysis

PHA: Phonetic Analysis (deals with the sounds of speech and their production, combination, description, and representation by written symbols.)

POA: Phonological Analysis(deals with the study 0(speech sounds in a language with reference to their distribution and patterning and to tacit rules governing pronunciation.)

MOA: Morphological Analysis(Deals with root / base form of the word and the morphemes affixed to it. The morphological analysis of a word in English can be shown as follows:

Indecipherability: $[_N[_A\text{in}[_A[_V\text{de}[_N\text{cipher}]]\text{able}]]\text{ity}]$.

SYA: Syntactic Analysis (It deals with grammatical analysis of sentences or discourse structure.)

SEA: Semantic Analysis(Concept-based analysis.)

DIA: Discourse Analysis(Analysis of the discourse structure by using knowledge of the world)

Corpora

CP:Text Corpora (A collection of writings used for linguistic analysis e.g. British National Corpus.)

SCP: Speech Corpora (A collection of recorded remarks used for linguistic analysis.)

Retriever

SEN: Search Engine(lt is a tool which searches for the required item from a given database.)

Recognizer

HRE: Handwriting Recognizer (It is a tool which performs automatic recognition of handwritten characters present in an optically scanned image.)

SRE: Speech Recognizer (Speech Recognition is the process by which a computer maps an acoustic speech signal to text.)

OCR: Optical Character Recognizer (It is a tool which performs automatic recognition of the characters present in an optically scanned image.)

SDT: Speaker Determiner (Recognition of the speaker from the speech s/he produces. A computer maps an acoustic speech signal and determines the speech characteristics of the speaker.)

Analyzer

SUT: Super Tagger

POT: POS Tagger (Part of Speech tagger.)

SET: Sense Tagger

MOP: Morphological Parser (Morphological Parser -Describes (a word) by stating its part of speech, form, and syntactic relationships in a sentence.)

SVP: Syntactic Parser (An algorithm or program to determine the syntactic structure of a sentence or string of symbols in a language.)

SEP: Semantic Parser (Gives a semantic structure or meaning representation in the form of a parse tree or parse tree fragments.)

WOS: Word Segmenter (Identifies discourse structure as a combination of words.)

SES: Sentence Segmenter(ldentifies discourse structure as a combination of sentences.)

WOP: Word Predictor (A tool that predicts the combination of certain characters as a word.)

PHP: Phrase Predictor (A tool that predicts the combination of certain character as a phrase.)

TST: Text Structure Analyzer

DAR: Discourse Analyzer (Analyzes a connected representation of the text by linking different descriptions of the same entity in different parts of the text.)

LEA: Lexical Analyzer

GRC: Grammar Checker (It is a tool which checks the grammar of the given sentences / discourse structure.)

SUM: Summarizer (A tool which summarizes a given text automatically.)

SPC: Spell Checker(It is a tool that checks the spelling of the given words with the help of a dictionary.)

SEL: Semantic Analyzer (Analyzes the input and gives a semantic structure or meaning representation or logical form.)

MOR: Morpher

STM: Stemmer (A tool which identifies character groupings which constitute the basic of text, such as root or stems of words.)

TKN: Tokenizer (Tokenisation is the identification of character groupings which constitute the basic units of text, such as words and punctuation. E.g. .he didn ‘t arrive ‘.

can be tokenized as:

<w>he</w><w>did</w><w>n ‘t</w><w>arrive</w><c> .</c>)

IND: Indexer

Generator

COD: Content determiner (Content determination involves decisions regarding the information which should be conveyed to the user.)

DIP: Discourse Planner (Even though text planning specifies the text down to clausal units, it does not always address the problem of organizing these clauses into sentences and sentences into discourse. Discourse planner takes care of that. The basic goal is to map conceptual structure onto linguistic ones, e.g. selection of a proper syntactic structure referring expressions, and content words (lexical choice).

SPS: Speech Synthesizer (Speech synthesis is the task of transforming written input to spoken output. The input can either be provided in a graphemic/ orthographic or a phonemic script, depending on its source.)

SUR: Surface Realizer (generates the individual sentences in a grammatically correct manner, e.g. agreement, morphology.)

LEX: Lexicalizer (Generates the words in morphologically correct manner)

Solution Number: Several solutions can be generated for each resource. Each solution may be identified by a separate (two digit) number.

Version Number: Each solution may be modified from time to time based on feedbacks received from users or after testing. Each modification may be identified by the version (a two digit) number.

Document Type: The solution for any resource may be documented through several types of documents. The types of documents can include

OVD: Overview document

RSD: Requirement specification document

DED: Design document

TSD: Test specification document

TRD: Test results document

PUB: Publication

REP: Report

RED: Resource Document

STD: Standards

SFW: Software

(More types can be identified and corresponding three letter codes be proposed)

Page N umber: Each document will have several pages and each page may be identified by a number

The footer will consist of

Name of the author(s) of the document

Organization

Month of creation

Whenever a field is not relevant with regard to a document a dash (-) may be added to that field

Samples:

Page 3 of second versions of two different requirement specification documents on speech synthesizer for Oriya language:

OR	PR	GEN	SPS	01	02	RSD	3
OR	PR	GEN	SPS	02	02	RSD	3

Page 4 of the second version of the resource document on Kannada fonts

KN	ER	OUT	FNT	01	02	RED	4

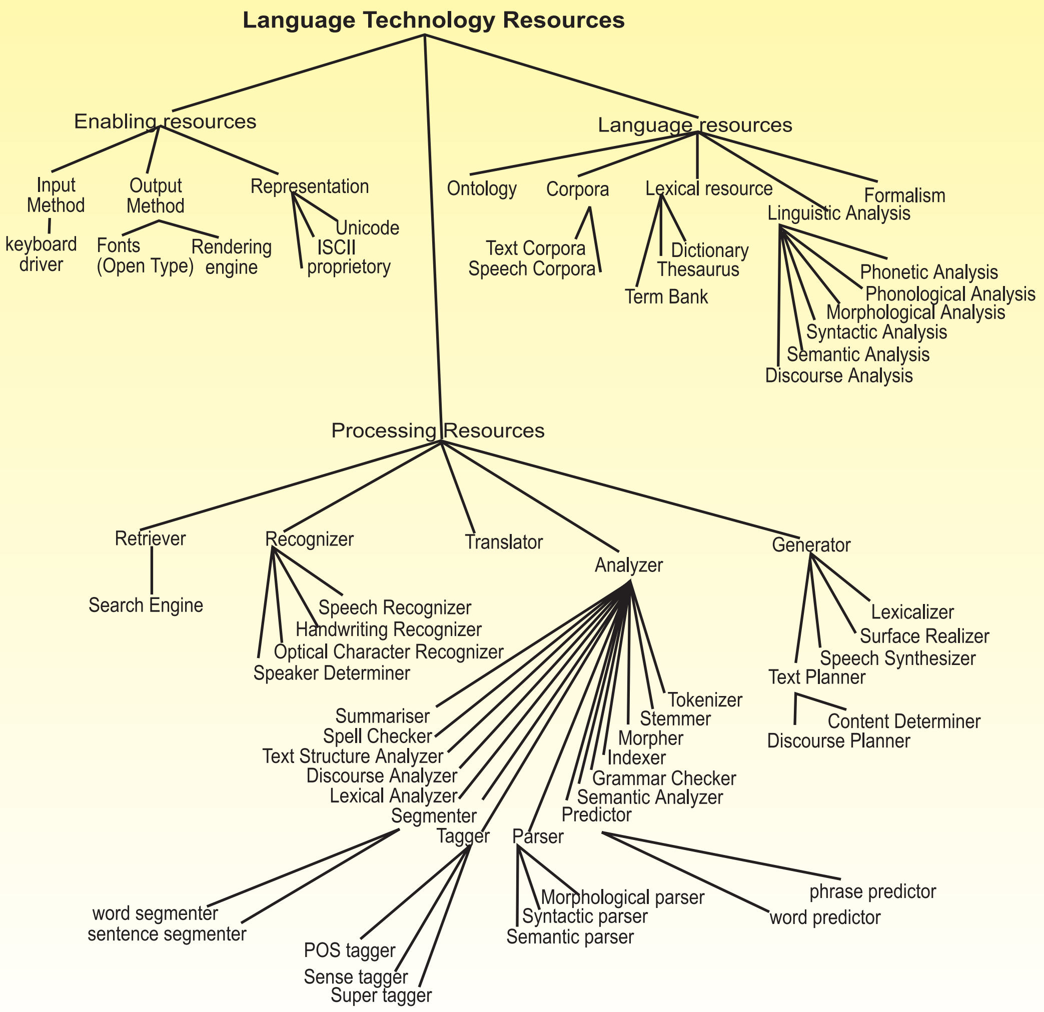

Figure 1. Language Technology Resources

(Courtesy: Prof. N.J Rao,
Ms. Kalyanmalini Sahoo
Centre for Electronic Design & Technology (CEDT)
Indian Institute of Science,
Bangalore-560012
Tel. : 080-23092377
E-mail : njrao@mgmt.iisc.ernet.in)