TAGSET Converter Documentation

LTRC , IIIT Hyderabad

May 9, 2012

Contents

List of Tables

1 Introduction

This tool aims at conversion of Part-of-Speech(POS) Tags from one scheme to another. Conversions can be of different types : single mappings , multiple mappings or context dependent conversions etc.

2 Running the Tool

2.1 Arguments

Option	Details
-i	Folder with Input Files
-o	Folder for Generated Output Files
-m	Map File
-w	Folder with Word-Lists

Table 1: Mandatory arguments and their Description

2.2 Run

```
>> ./extract.py -i inputFolder -o outputFolder -m mapFile -w wordListFolder
```

2.3 Sample Run

```
>> ./extract.py -i dataset/test/ -o dataset/new_files/ -m map.xml -w wordlist/
BIS_wordlist/
```

3 Description of Files

4 Customization

For customizing the Tool for any generic tagset conversion or for POS Tag conversion for languages that have different tagsets than hindi, the user needs to change the Map File. A sample *map.xml* is given with the Tool. User can also put up respective Word-Lists for tags that have one-to-many mappings.

Example : For Rule, $CC \rightarrow CC\_CCD, CC\_CCS$
User needs to create 2 Word-Lists , each for CC_CCD and CC_CCS. Failure to do will result in the tool assigning the first tag from the multiple destination tags specified in the Map File.

File	Details
extract.py	Source Code.
argumentParser.py	File to parse the command-line arguments given with the main code.
map.xml	Contains the TAG mapping rules. Can be changed.
wordList	A sample folder containing Word-Lists for BIS TAGS.
–wordList/BIS_wordList	Folder that contains the BIS_Tagset Word-Lists.
–wordList/Other_wordList	Folder to put Word-Lists for other Tagsets.
dataset	Folder with sample Files for conversion.
–dataset/test_input	Folder with sample Input Files.
–dataset/test_output	Folder with sample Output Files.
–dataset/test_gold	Folder with sample Gold Files.

Table 2: Files in the Package and their Descriptions

4.1 Map File

The Map File is to provide the mappings from Source Tagset to Target Tagset (Referred at many instances as Destination Tagset). A sample Map File is given in Listing 1 .

```
<scheme>
        <rule>
                <sourcetag>RDP</sourcetag>
                <expression>{T2-T1-(RDP)=>T2-T1-T2}</expression>
        </rule>
        <rule>
                <sourcetag>NNP</sourcetag>
                <destinationtag>N_NNP</destinationtag>
        </rule>
        <rule>
                <sourcetag>CC</sourcetag>
                <destinationtag>CC_CCD</destinationtag>
                <destinationtag>CC_CCS</destinationtag>
        </rule>
</scheme>
```

Listing 1: A Sample Map File

A Description of XML Tags used in creation of Map File is given in Table 3.

4.2 Type of Mappings

Mappings are conversions from a set of tags to another set of tags. Following are the types of mappings and how to handle them in Map File :

Tag	Details
<scheme>	This is the root node hovering over the rules.
<rule>	Node for describing a single mapping rule.
<sourcetag>	The POS Tag on the left side of the rule.
<destinationtag>	The POS Tag on the Right side of the rule. Each POS Tag needs to be specified independently.
<expression>	For Context dependent Rules (Ex. RDP-Reduplication), user can write expressions [see Section 4.4]. These expressions are captured in this node.

Table 3: XML Nodes and their Specifications

1. **One to One** : LHS and RHS, both have single Tag entry.

   Example : $NNP \rightarrow N\_NNP$
   This requires a single rule with single sourcetag Node and single destinationtag Node.

2. **One to Many** : LHS has a single Tag entry, while RHS has multiple Tag entries.

   Example : $CC \rightarrow CC\_CCS, CC\_CCD$
   This requires single rule with single sourcetag node and multiple destinationtag Nodes. To know how one to many mappings are resolved see Section 4.3.

3. **Many to One** : LHS has multiple Tag entries, while RHS has a single Tag entry.

   Example : $CC\_CCS, CC\_CCD \rightarrow CC$
   This requires multiple rules with single sourcetag Node and single destinationtag Node. The multiple entries in the LHS make up for different rules in the map file.

4. **Context Sensitive Rules** : A special case in which the RHS is dependent on neighbouring elements in the sentence. Expressions provide a way to incorporate the context into the mapping rule. Listing 1 shows the example for RDP-Reduplication. To know how they work, see Section 4.4

4.3 Word Lists

Word Lists are provided to convert from coarser Tags in Source TagSet to finer Tags in the Target TagSet. This can be easily done for Tags with a closed class of lexicon, such as pronouns, demonstratives etc. Word Lists are crucial for implementation of Many to One mapping. Each Tag in the RHS needs to have a seperate File with words belonging to that particular category. The name of the file with Word List is same as the Tag. Example : $CC \rightarrow CC\_CCS, CC\_CCD$
For the above rule, 2 Files will be created, namely CC_CCS and CC_CCD. CC_CCS will contain words that belong to Subordinating Conjunction category and CC_CCD will contain words belonging to Co-ordinating Conjunction category.

4.4 Expressions

Expressions are rules to handle context sensitive mappings. An expression is of the form : $A \rightarrow B$, where A is the LHS, depicting the current expression to be searched into the POS Tag File.

Example : **T1-T2-(RDP)** depicts an expression of length 3 with T1, T2 as Tags at 1st and 2nd Positions respectively and "RDP" as the Tag at 3rd Position. Generic Tag names can be any alphanumeric combination, Ex. T1,T2 etc. POS Tags must be specified within rounded brackets, Ex. (RDP)

The RHS of an expression is the desired form in the Target File. The variables (Generic Tag Names) that have been assigned in LHS of the rule can be used in RHS to depict their new position.

Example : **{T2-T1-(RDP)=>T2-T1-T2}**

The above rule ignores the Tag Value for *T1* and *T2*, but changes the value of *RDP* to *T2*. This rule handles the Reduplication in Hindi, which is seperated by a "Hyphen" (*T1*) in most of the cases.