Multi-lingual Lexical Database and Corpus Management Tools

CDAC Noida and IIT Kanpur

Abstract

Machine translation system requires a lexical resource for the target language. AnglaBharati is rule-base machine translation system for English to Indian language translation and has well defined English-Hindi lex resources. While adapting AnglaBharati engine for other Indian languages like-Urdu, Punjabi, Bangla and Malayalam translation, it was felt necessary to have lexical resources for each of these languages. However it was necessary to maintain the consistency in these lexical resources. Thus the process was Normalization was designed. It is a process through which differences in lex resources at word and POS level can be identified automatically. The process of normalization leads to correction of lex resources as well. This paper describes the complete process of normalization and also various tools that have been developed to carry out the task of Normalization. It also describes the functionality of various tools developed for maintaining the lexical resources and managing the corpus.

Index Terms: Normalization, Lexical database, Lexicon tools, corpora tools

I. Introduction

Normalization can be defined as "A process, automatic or semi-automatic, for building a multilingual lex resource with appropriate matching of words, duly categorized into appropriate sub categories, with their linguistic information such as POS, GNP, inflections, semantic and paradigm". Sub categories can further be standardized based on linguistic knowledge, for validating the process following categorization has been worked out in such a manner that it does not stop user to add or modify the contents, since the first step of process is to create a superset of single language words, duly aligned with their linguistic information. After the superset creation, the differences can be traced and resolved. Normalization process can be carried out with sub categorized files. These files can be normalized based on words, POS, paradigm number used in files. Normalization can also be done for range of words, synonyms, equivalence etc. The rest of the paper is organized as follows: Section-II describes the complete Normalization Process, algorithms or tools used in Normalization. The implementation and results for the process are given in section III and section-IV discusses the various tools developed for lexicon and corpus management and section V concludes the paper.

II. Normalization Process

The Normalization process is not one time activity, but is a repetitive process and needs to be carried out after specific time periods.

The process is shown in Fig.1. The steps involved in the process of normalization are as follows:

- collection of lexicon resources of different languages
- Extraction of source language specific data
- Superset Creation

- Similarity & Difference identification
- Neutralizing the differences
- Updation of Lexical Resource

Extraction of source language specific data

The algorithm used for extraction of input specific data reads the nature of data whether it is Lexicon structured file or simple one to one mapping file and based on this identification it extracts the data. The structures of data file are as per the requirements of MT system and engine [3]. For AnglaMT system it can be done in two ways. For lexicon structured data, it extracts English word, part of speech with category code as the source language is English. For plain data it extracts the words only. The algorithm runs for the block of word separated by *** and then extracts the source language specific data and keep it at a separate location [3]. The process is repeated for all the data files.

Superset Creation

Superset is the data set that is to be achieved by all the sub systems. To create superset, data of source language from each of the sub systems which is kept in the repository has to be taken and merged and thee the repetition in the data is eliminated. The generated data is the required superset.

Similarity and Difference Identification

After acquiring the superset, difference has to be generated for each of the sub systems. For this an automated tool called "word list break" is used. It uses superset and data obtained by different subsystems to generate the differences for each of the subsystems. The difference data is extracted and shown in two forms viz. data to be added and data to be deleted. Similarly on comparison the common or similar data can be extracted.

Neutralizing the differences & Updation of Lexical Resources

In order to neutralize the differences, difference data is taken into consideration and the differences are incorporated in all the sub systems. This step uses the lexicon entry module. Once the set is ready the lexical resource is updated for each language.

III. Implementation

The Normalization process not only normalizes the data as per the superset but also increases the efficiency of the system. This approach has been implemented in the AnglaMT system for all the languages. The differences turned out to be huge in the initial iterations which reduced to quite an extent after successive iterations. The differences reduced as the iterations increased. For a set of 50000 words with 5 subsystems it took 5 iterations to get the minimal or nil difference. Currently there is no difference between the lexical resources of the different languages and all the systems behave in a consistent manner for a given sentence. Figure 1 shows the normalization process.

IV. Lexicon & Corpus Management Tools

There are large number of tools that have come up during the development of the AnglaUrdu and AnglaPunjabi systems. The tools could be related to lexicon, corpus or testing. We shall

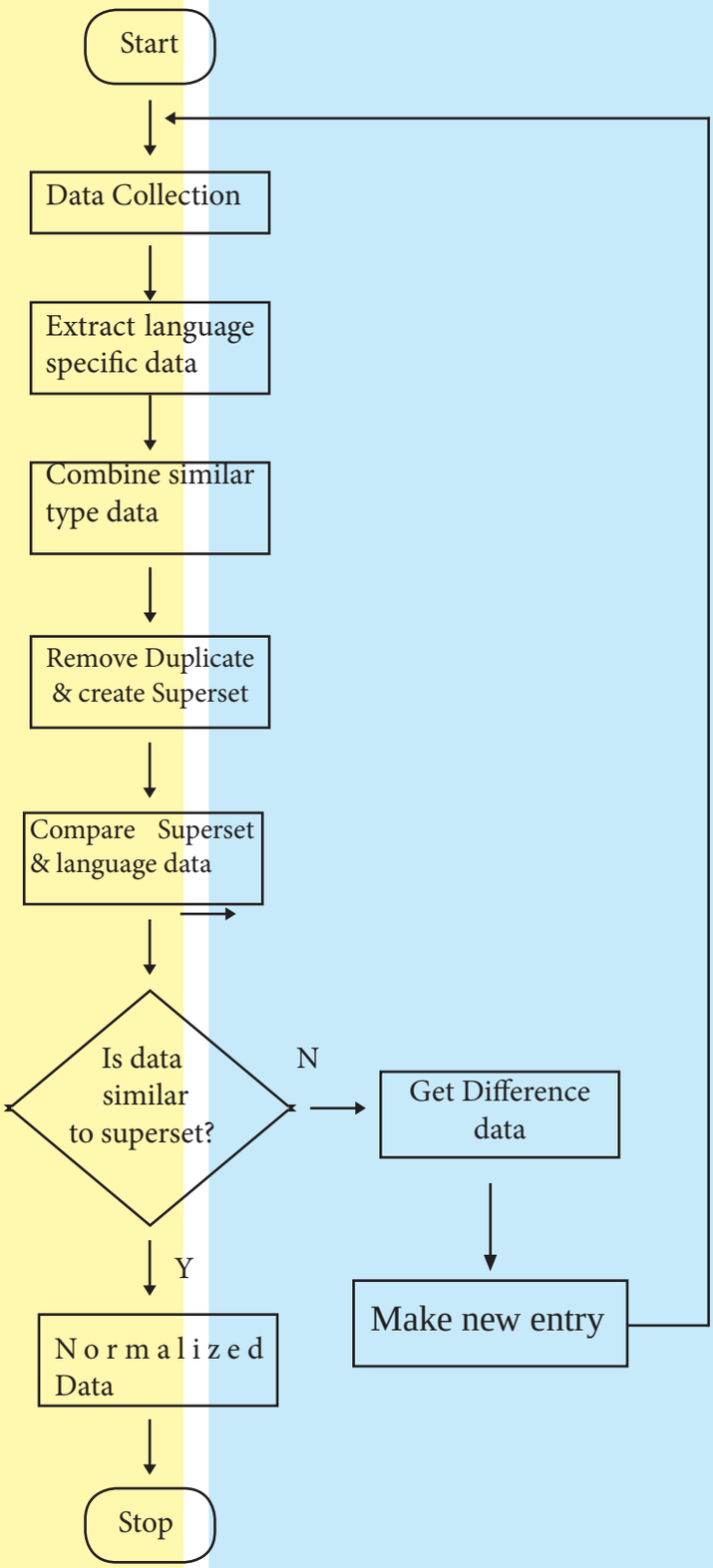


Fig 1: Normalization Process

give a brief functionality of each of these tools. The lexicon related tools deal with verification of lexicon i.e. whether the lexicon is in proper format or not and Lexicon entry interface, that helps in creation of lexicon etc. Similarly corpus related tools mostly deal with cleaning of corpus or extraction of relevant information from the corpus. Two broad categories of tools that have been developed are:

- Lexicon Related
- Corpus Related

Lexicon Related Tools

Lexicon Validator

The purpose of Lexicon validator is to ease the process of lexicon validation. As lexicons have specific structure, it helps to validate that structure. Lexicon has huge amount of data and it is very difficult to manually check the correctness of Lexicon and hence this tool helps in making this process user friendly and error free. It also reduces the effort and time taken to perform the validation. The Lexicon validator has the following features:

- It performs structural corrections.
- It performs logical corrections which imply corrections for dependencies as per the entries in lexicon.
- It can generate the tabular (excel sheet) view of lexicon.
- It can generate lexicon from the excel sheet that represent lexicon.
- It can extract parts of speech dependent lexicon from a lexicon.

One can use the "Select File" button to select the file for which validations are to be performed. Then buttons "I.Correct and S.Correct" can be used to check and finally correct the file. In case of any errors system messages will be displayed. Fig. 2 shows the lexicon validator.

An xls file can be created from an existing text lex file and vice-versa by using "lex to xls" or "xls to lex" buttons. Also specific part of lexicon based on Part of Speech can also be extracted

by selecting the lexicon category and using the

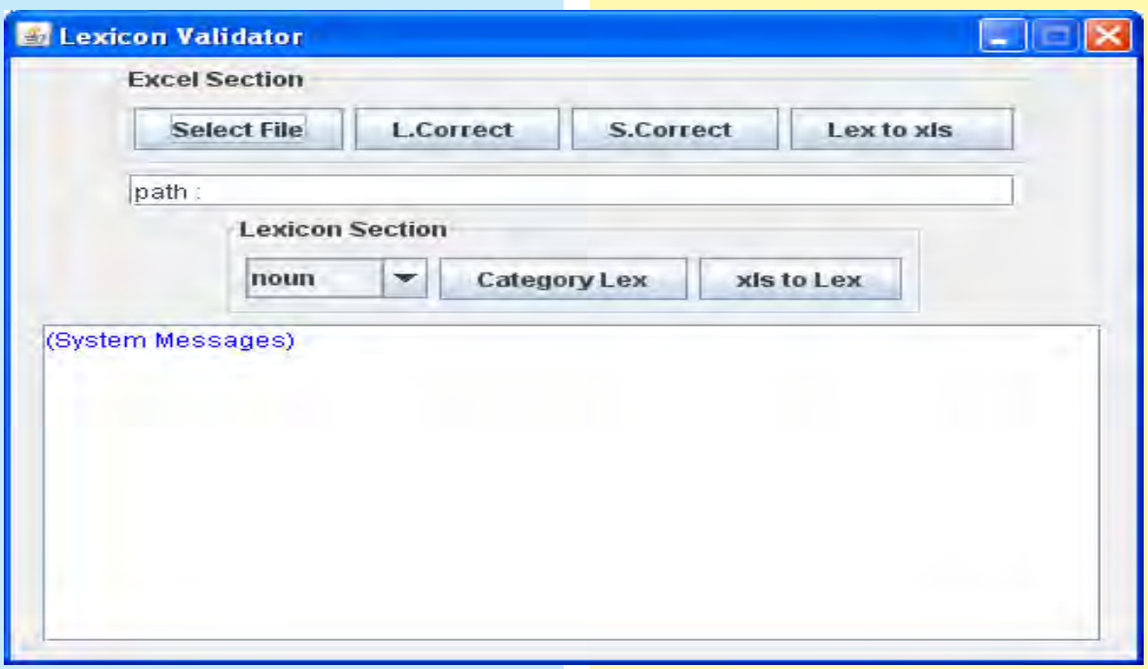


Fig. 2 Lexicon Validator

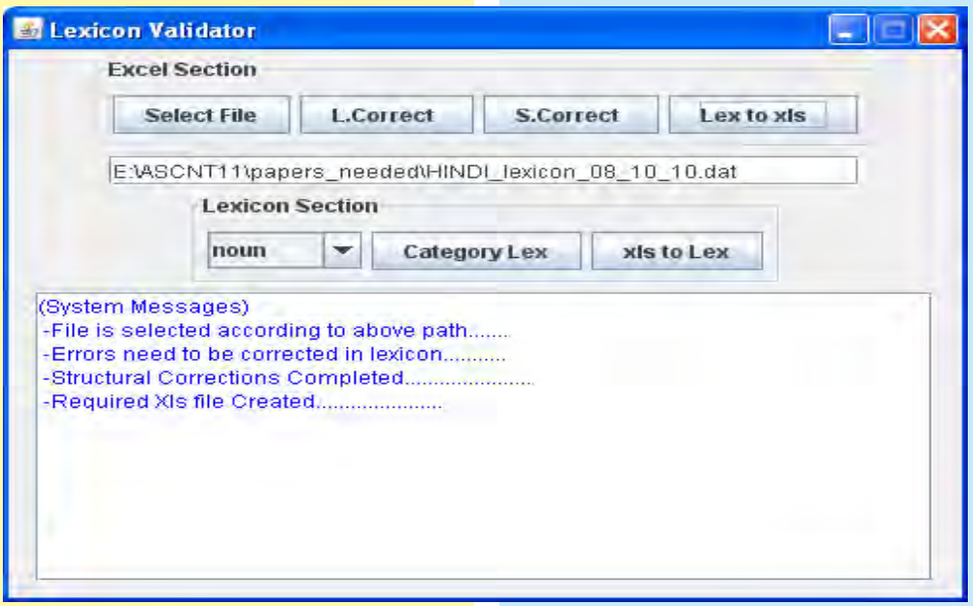


Fig. 3 xls creation

"Category Lex" button. Fig. 3 shows the process.

Word Counter

The purpose of this tool as shown in Fig. 4 is to count the words entered in lexicon. It helps in evaluating the number of entries made by the entry operator in the lexicon. You can select the lexicon file for which you wish to count the words and use "Count" button to perform counting.

Lexicon Manager

Lexicon Manager shown in Fig. 5 is a tool that has been designed to extract information regarding the POS, Semantic Tag, Category Code, Paradigm number and the meanings of a specific word already available in the lexicon. This is used to view, add, modify and update the entries currently present in the lexicon.

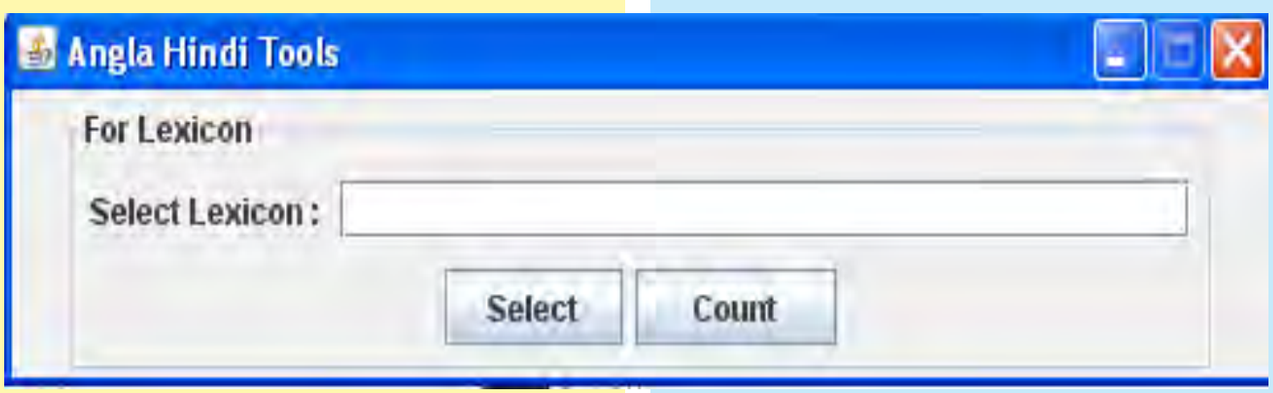


Fig. 4 Word counter

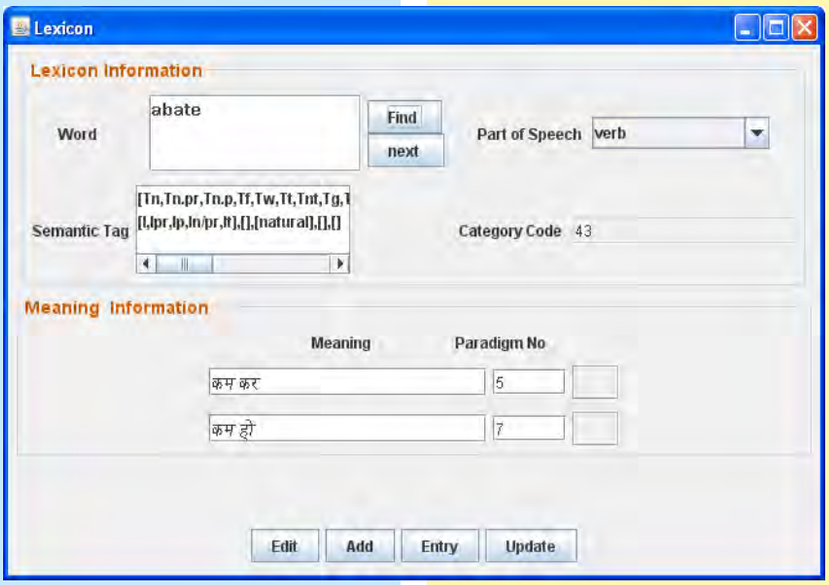


Fig. 5 Lexicon Manager

Lexicon Creator

The Lexicon Creator tool (Fig. 6) helps in the lexicon creation. Using this tool we can generate lexicons as per specification. Since the task is automated, there are minimal chances of any errors being introduced in the lexicon creation. If we perform the same task manually then there are chances of introducing some errors due to incorrect structure that may appear unknowingly.

All the options are provided in the interface. All the possible semantic tags available have been provided and the user can select the suitable options. Also the POS option, category code etc. are given and these can be selected as needed. The available semantic tags for human are shown in the Fig. 7.

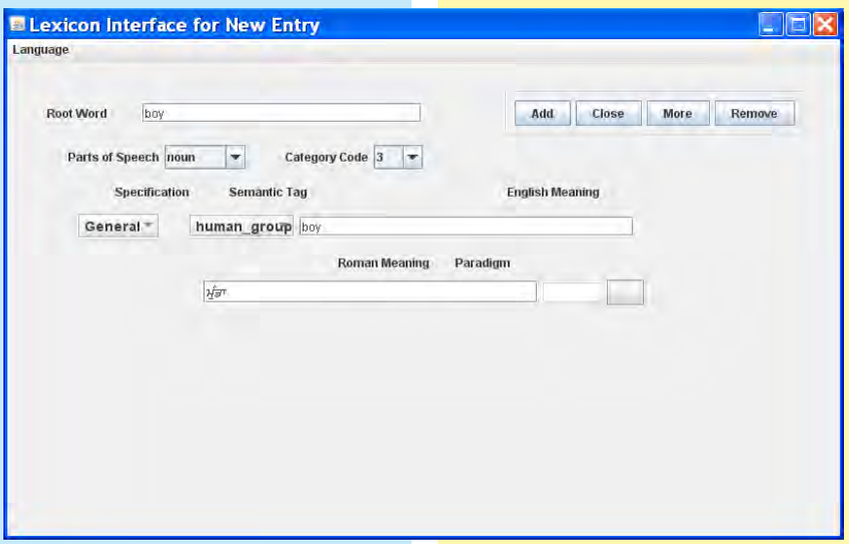


Fig. 6 Lexicon Creator

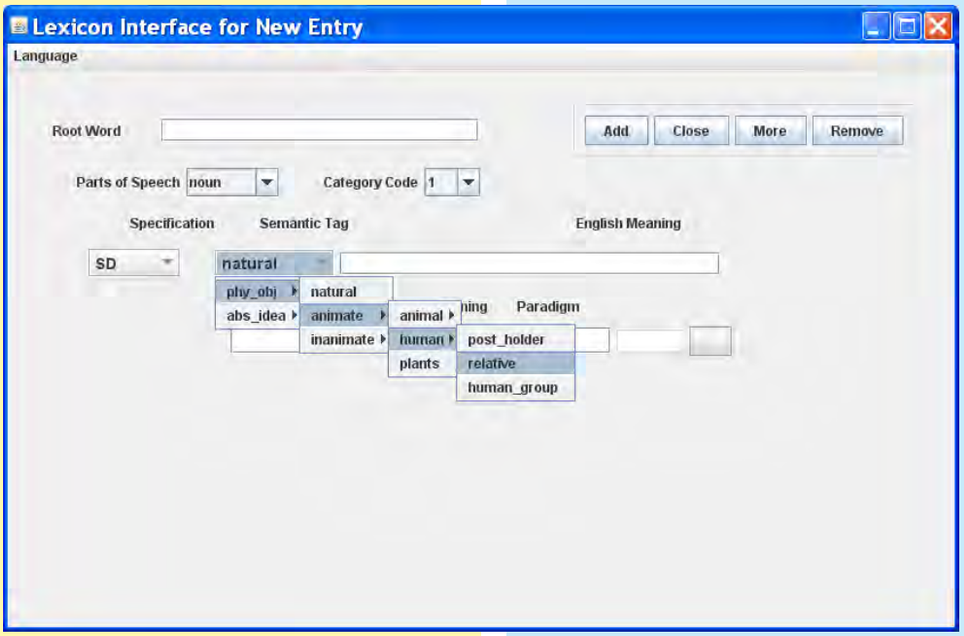


Fig. 7 Semantic Categories

Lexicon Normalization Tool

The normalization tool is shown in Fig. 8. This is a very powerful and important tool as it helps in maintaining the consistency of the Lexicon being updated, modified etc. by the different members of the consortium independently for their own languages at their centers.

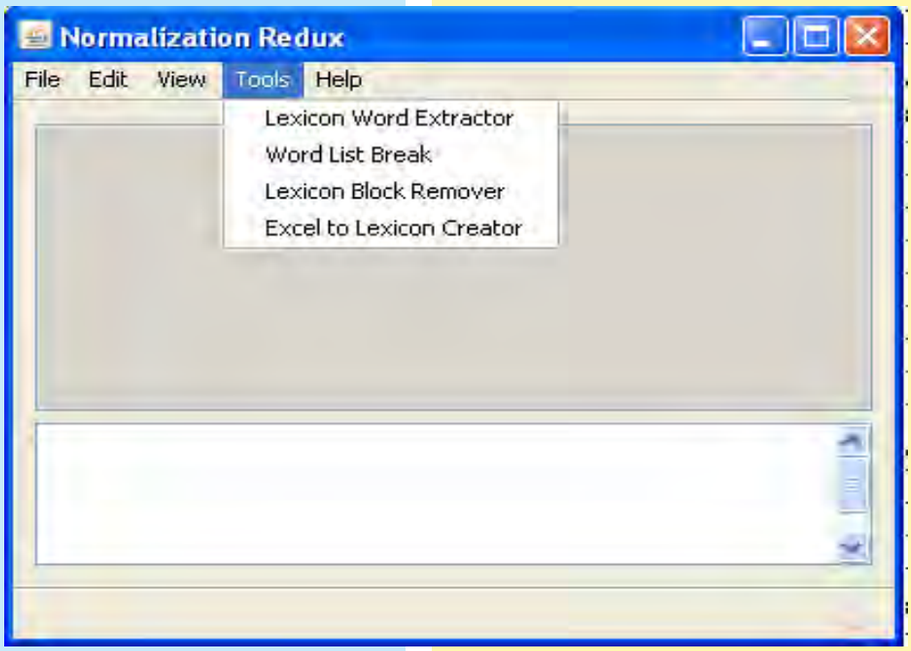


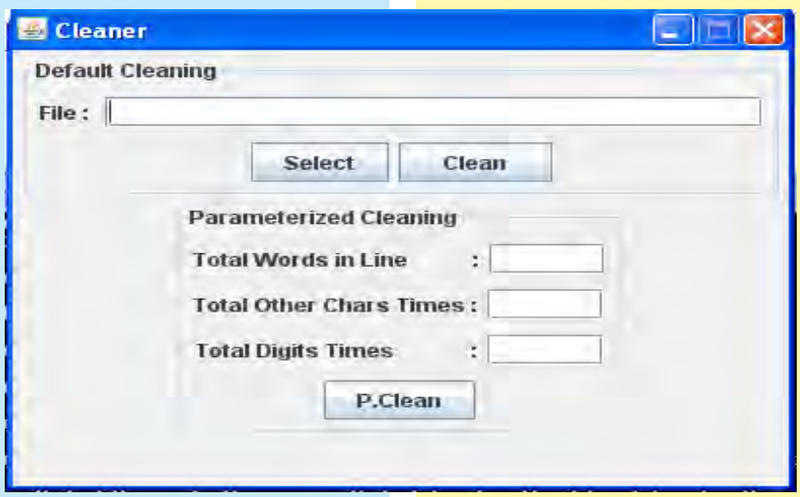
Fig. 8 Normalization Tool

To make sure that the system behaves in the similar manner for all the languages, this is one component (lexicon) of the system that should remain consistent for all the languages and thus after a specific time period Lexicon for all the respective centers are compared to find out inconsistencies, if any. This is a recursive task and being carried out weekly.

Corpus Related tools

Corpus Cleaner

This tool (Fig. 9) is used to clean the raw data. It is a semi automated tool that can clean the data to approximately 90 %. It reduces the manual work to quite an extent. It has two options for cleaning the data namely default cleaning and parameterized cleaning.



One can use either of the two given options for cleaning. However one needs to specify the parameters like total number of words in a line or presence of some special characters in the text etc.

Fig. 9 Corpus Cleaner

Sentence Boundary Program

It is one of the optimize d semi automated tools developed during the development process of AnglaUrdu and AnglaPunjabi systems (Fig. 10). It is developed to ease the manual efforts related in marking the sentence boundaries for sentences available in clean corpora. It has shown accuracy of 95%. It marks boundaries by using the symbol <s>……</s> based on the criteria mentioned on the basis of which characters are to be taken as the sentence terminators.

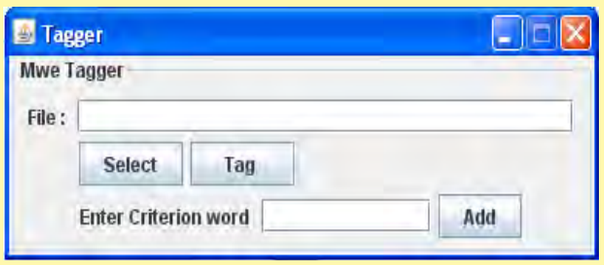


Fig. 10 Sentence Boundary Program

Multiword Expression Extractor

This tool (Fig. 11) is used to generate arbitrary list of multi words from the tagged output of the Charniak parser. It is based on the rules provided for the extraction. This tool has multiple functions. It is also used to extract the list of words which are candidates for unique words, Named entities, Noun phrase, Verb Phrase, Adverb Phrase and several MWE.

Fig. 11 MWE Extractor

V. Conclsion

Though normalization approach is general in nature but currently it has been implemented for AnglaMT systems. By making some modification in its general structure one can use this approach in any MT system. The modification would only require the change in format or structure of data to be mentioned as the data structures may vary from system to system. After Normalizing, these lexical resources can be used further to generate several files such as MWE_lex containing multiword expressions [1]. The tools developed during all the process have reduced the possibility of number of errors being introduced in the manual process and also the time taken to perform these tasks manually.