AnglaUrdu: English to Urdu MT system based on AnglaMT Paradigm

Abstract

This paper presents an approach to derive English to Urdu Machine Translation, by adding / modifying components to AnglaBharati system designed by IIT Kanpur for English to Hindi Translation. Since Urdu and Hindi are culturally similar languages, change or modification in RuleBase is not required. Although, Urdu has multiple origin (Pasto Arabian, Indian etc.), only spoken Urdu in Jammu Kashmir, Hyderabad, Lucknow & nearby area has been considered, so as to maintain similarities between Hindi & Urdu. This paper shows need, format & creation of "ITable", to be included and to get near perfect Urdu translation without much modification in architecture of the engine. The formats of ITable have been designed suitably based on the study of similarities and dissimilarities between Hindi & Urdu. The "ITable" directly guides post processor to form appropriately part of sentences, which otherwise gets generated in ill-formed structure by the engine. Thus, "ITble" not only guides translation process but also helps post processor, which is based on the cost spent on translation. Structure of "ITable" matches with not much change in lexical resource except for a pointer field to be used for search in ITable. It can be updated with ease without major modification for getting the higher version of Machine Translation.

I Introduction

Hindi/Urdu or Hindustani is the most widely spoken language on the Indian subcontinent and India's official language alongside English. It is an Indo-European language, being a distant relative of common European tongues such as English, French, Russian etc. Hindi and Urdu can be considered as two complementary sister languages, where Hindi in its higher registers is deliberately Sanskritized, while Urdu prefers to use words borrowed from Arabic and Persian, giving it a more "Muslim" flavor. Hindus will often say they speak Hindi, while Muslims will often say they speak Urdu. However, as spoken on the street, Hindi and Urdu have a high degree of mutual intelligibility .

Hindi and Urdu are sister languages having common origin (Hock, 1991). They are structurally very close to each other and use similar postpositions, verb morphology as well as complex predicate verb structure. We have exploited this similarity in developing English-Urdu translation from the English-Hindi machine translation (MT) system.

System's ability to capture adequate knowledge from the source language so as to be able to generate the target language text that is truly a translation of the source language, decides both the quality and accuracy of translation. A compromise is to use a pseudo-interlingua (Sinha, 2004) for a class of structurally similar languages. Here a text generator for each of the target language needs to be developed. In this paper, we present a more straightforward strategy for deriving translation to Urdu from Hindi English-Hindi MT system without the need for developing a detailed grammatical analysis of the source language and without developing a full fledged text generator from the Interlingua representation. We primarily use lexical mapping of Hindi words to Urdu and transform the output sentence for gender agreement in case of dissimilarity in gender of the lexicons. It should be noted here that there is no correlation in the scripts in which Hindi and Urdu are written. Hindi is written in Devanagari which is written left to right. Urdu script is based on Perso-Arabic alphabet with six additional characters primarily to map sounds of English and Hindi.

II Modules Involved in AnglaBharati Translation

It is an overview of the English to Urdu Transltion system, which is developed on the

basis of existing approach used for AnglaHindi RBMT. Its beta-version has been made available on the internet for free translation at http://tdil-dc.in. AnglaUrdu is an English to Urdu version of the AnglaBharati translation methodology developed by CDAC Noida. Anglabharati is a pseudo-interlingual rule-based translation methodology. The primary modules are detailed as under.

1. User Interface:

i) Desktop version: It is the base domain specific interface developed on linux platform.

ii) Web version: The web version is based on domain categorization and the an evaluation mechanism is provided with it.
UI(both) versions are provided with the functionality of Roman to Unicode conversion .

2. Preprocessor/Postprocessor: It contains the information for the modules used by the preprocessor to retrieve various entities from the text, which can be, treated as separate entities before sending the text to the Translation engine. It also gives the information about the Postprocessor, which reassigns the code values. There is not much change in this module for developing Eng-Urdu MT system.

3. Example Base:

It contains the information about the translation of the sentence by matching it with the stored patterns.

4. Rule Base:
It contains the information of the rules for parsing sentences. This is the main translation engine, which identifies the structure of English sentence and transforms it into an intermediate form which is PLIL(Pseudo lingua for Indian languages)

5. Morphological Analyzer:
It contains the information about the files used in the Morphological Analyzer (morph). Morph searches a given word or groups of word in the lexicon and generates the files in the form of prolog files for Rule Base. The current lexicon used is
57 K.

6. Hindi Text Generator:
This contains the information of the files used in the text generator module. Textgenerator converts the pseudo code given by the earlier stages of the system into the target language.

III ITable translation mechanism for Urdu Language

It shows that for translation purpose we first translate the English sentence in to Hindi through the rule based existing Rule Based system, and PLIL (pseudo lingua for Indian Languages) is generated. A rule base is used for this transformation. The PLIL representation is then transformed to the target language using a text generator. The English-Hindi MT system produces a Hindi translation. All grammatical analysis of English is performed by this translation engine. The parse structure of the input English sentence is transformed to the corresponding Hindi structure. The output Hindi sentence is generated from this structure.
English to Hindi translation examples:
Ex. PLIL He is going to the market.

<aff {sub_np (he noun masculine singular third [human] [vaha:m 7] [] []) } {pp (the det [] [anda] [A]) (market noun neuter singular third [place] [bAjZAra:m 7] [] []) (to prep [to]) } {main_vp_active (go verb_5 normal is masculine singular third [jA] 3 [] []) } > . sviram

Though Hindi and Urdu sentence structure are

same, the individual words may differ in gender, number and may have multiple parts of speech having different meanings and POS Therefore, in order to derive the correct form of the Urdu sentence from Hindi with number and gender agreement, the English-Hindi MT system must also produce the POS, gender and number information for each of the Hindi word or word groups produced by the translation system.

New Module Developed for AnglaUrdu Translation

Urdu Text Generator

1.1 Automated Lexical Expansion

1) Auto generation of ITable (Morphological Info): Expanding the readily available lex resource with the Hindi morph information. This information is easily available in existing RBMT engine.

2) Format for ITable

The Changes in Number and Gender with no changes in Person are as follows.

TABLE I: *Number*: S :: Singular, P :: Plural, D :: Direct

S.no	Root Meaning	POS	Num	H_G	U_G	U_ equivalent
1	paSuSAlA (पशुशाला)	noun	S & D	F	M	vLrcy
2	paSuSAlA eM (पशुशालाएं)	noun	P & D	F	M	vLrcyk
3	paSuSAlAoM (पशुशालाओं)	noun	P & O	F	M	vLrcyk̇
4	anuwwariwa (vuÙkjfr)	adjective	NC.	NG	NG	बेजताबी
5	upaxeSa (उपदेश)	sadjnoun (adjective + noun)	S & D	M	F	नसीहत
6	nasIhawa (नसीहत)	sadjnoun (adjective + noun)	P & D	M	F	नसीहतें
7	upaxeSoM (उपदेशों)	sadjnoun (adjective + noun)	P & O	M	F	नसीहतों
8	XvaniroXiwa (ध्वनिरोधति)	adjective+Verb4	NC.	NG	NG	गूँज को रोकने वाला
9	AacCAxana karanevAlA (vkPNknu djuokyk)	adjective+Verb5	NC.	NG	NG	ढकने वाला
10	Aur AXika (vkj vf/kd)	adverb	NC.	NG	NG	और ज्यादा

3) Map to Urdu equivalent adjective and adverb substitutions done when needed.

a. Predicate verb differences to be taken care of

b. dissimilar nouns , groups of nouns

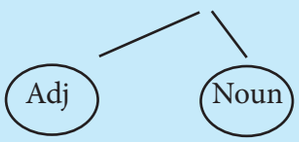
c. Adjective, adverb differences

4) Filtered the Inflection Itable from 4,00,000 to almost 1,00,000 entries

The Hindi entries having similar usage in Hindustani Urdu is reduced or deleted from ITable.

1.2 POS resolution

Hindi Translation- यह बात आम है



1.3 Gender change-

S.No.	Example Type	Urdu Root Word	Inflections
1	Prefix modification	किताब	कुतुब
2.	Prefix addition	इल्म	उलूम
3	Suffix addition	माहिर	माहिरीन
4	Total word change	दवा	अद्वियात

It is used to change the gender of the associated post positions Since the gender of the verb depends upon the gender of the subject or the object in both Hindi and in Urdu, it may also require changing gender of the verb.

Ex. mere Ane kA kAraNa

Hindi Translation: मेरे आने का कारण (object gender-Masculine)

Urdu Translation: मेरे आने की वजह (object)

(object gender-Feminine)

Post position is dependent on the object and the sentence nature will be changed accordingly from Hindi to Urdu

1.4 Transliteration for Hindl to Urdu is provided for unknowns literals.

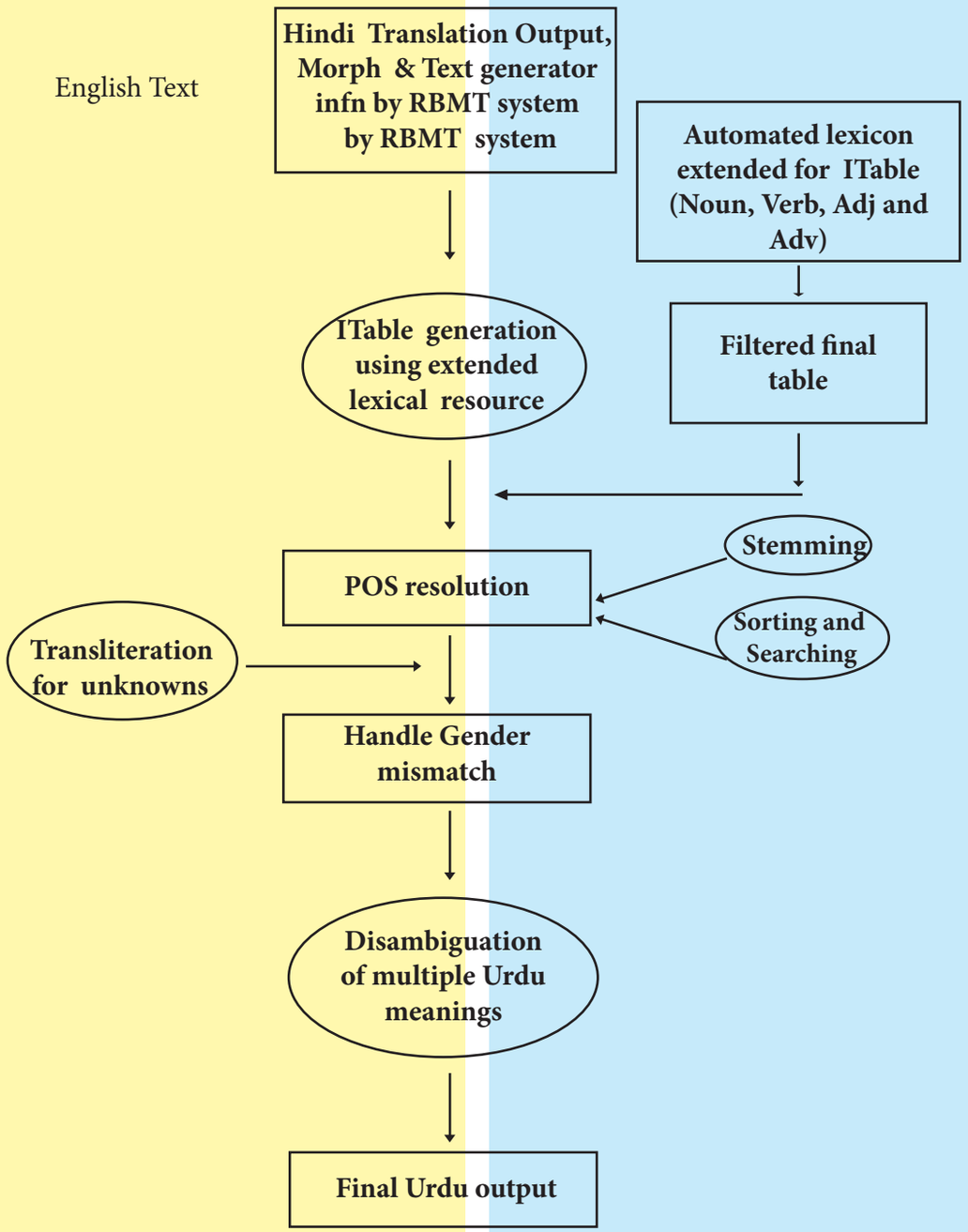
IV Reasons for designing ITable to Paradigm approach

The approach: A paradigm table is a set that shows the ways to conjugate a verb, decline or inflect a noun, etc., in all possible ways, by using a model (the word Paradigm means model or example). Paradigm number is a number which is assigned to root words that belong to different paradigm classes. The paradigm rule has been framed keeping in mind the inflections of a Urdu verb & noun etc.

In paradigm approach, for Urdu translation, a long list of different inflections used in Urdu language were generated, which is not exhaustive in nature. With time and proper linguistic analysis that list goes on increasing each day, which in turn posed some limitations on system implementation of the paradigm table. In Hindi a root word is inflected with the suffix addition only, but in Urdu language both suffix and prefix inclusion is carried out.

This limitation on the paradigm approach for Urdu language has shown the way to design a new approach for the translation.

AnglaUrdu Translation Process:



V Usage of ITable for Translation Purpose (Search Mechanism)

i. Translation output for Hindi language

ii. Exploit given morphological information

iii. Stemmer is used to terminate the plural and feminine suffixes before starting search in the ITable.

iv. Search for the longest string match.

v. Unknowns are transliterated in Urdu.

vi. Gender change module

VI Translation quality using Itable.

Englsh Sent:
I, who am your Prime Minister, will lead you.

Translation without using I.Table
mEM ApakA rAjA hUz Ora ApakI rahanumAI karUzgA
मैं आपका राजा हूँ और आपकी रहनुमाई d :rxk

Translation using I. Table
mEM ApakA vazIre Ajama hUz Ora Apaki rahanumAI karUzgA
मैं आपका वज़ीरे आजम हूँ और आपकी रहनुमाई d :rxk

Ex2.
Englsh Sent:
Some Ladies were the registered members of the association.
Translation without using I.Table
kuCa oraweM asosieSana kI meMbara xarja WIM
कुछ ओरतें असोसएिशन की दर्ज मेंबर थीं

Translation using I. Table
kuCa KavAwIna asosieSana kI meMbarAna xarja WIM
कुछ खवातीन असोसएिशन की दर्ज मेंबरान थीं

There are transliteration errors due to phonetic differences in the way the names are written in Hindi and Urdu. This is also due to one-to-many mappings of some of the Hindi consonants to Urdu consonants.
The words which are pronounced in a different manner and provide wrong transliteration are included with the correct pronunciation in the ITable.

Ex.
Terminology:
Words: pneumonic, psychiatric
Urdu Transliteration-> पU;rekfud, ilkbfd,fVd
Desired Transliteration-> U;rekfud lkbfd,fVd

Named Entity:
Words: Mecca Madina, Xavier
Urdu Transliteration-> eDdk मदीना, tkfovj
Desired Transliteration-> eDdk मदीना, tfovj

AnglaUrdu Interface:

The AnglaUrdu system aims to design, develop and deploy a Machine Translation (MT) System from English to Urdu Language in Tourism and Health Domains. As a result, two versions of interfaces were developed (i) web version and (ii) desktop version. This paper explains Desktop interface only.
Desktop GUI: This GUI has been designed for

users working on standalone PCs. After loading the system, an editor will be displayed (Fig.1). It contains many options for text editing and contains many user friendly menu items. In this GUI, user can input sentence(s) directly or open a file from the desired directory in the PC by using the "Open" menu item. Facility for English spell checking is also provided in the software. After inputting the sentence(s) or file for translation click on Translate menu item and sentences will be translated as shown in Fig.2. The resulting window will contain these sentences arranged in grid format . Fig.2 shows the translation window where all sentences will have their corresponding translations along with multiple alternatives, if any. From the multiple alternatives user will select the best possible sentence and still if not satisfied with the output can do the correction using Punjabi keyboard as shown in Fig.3.

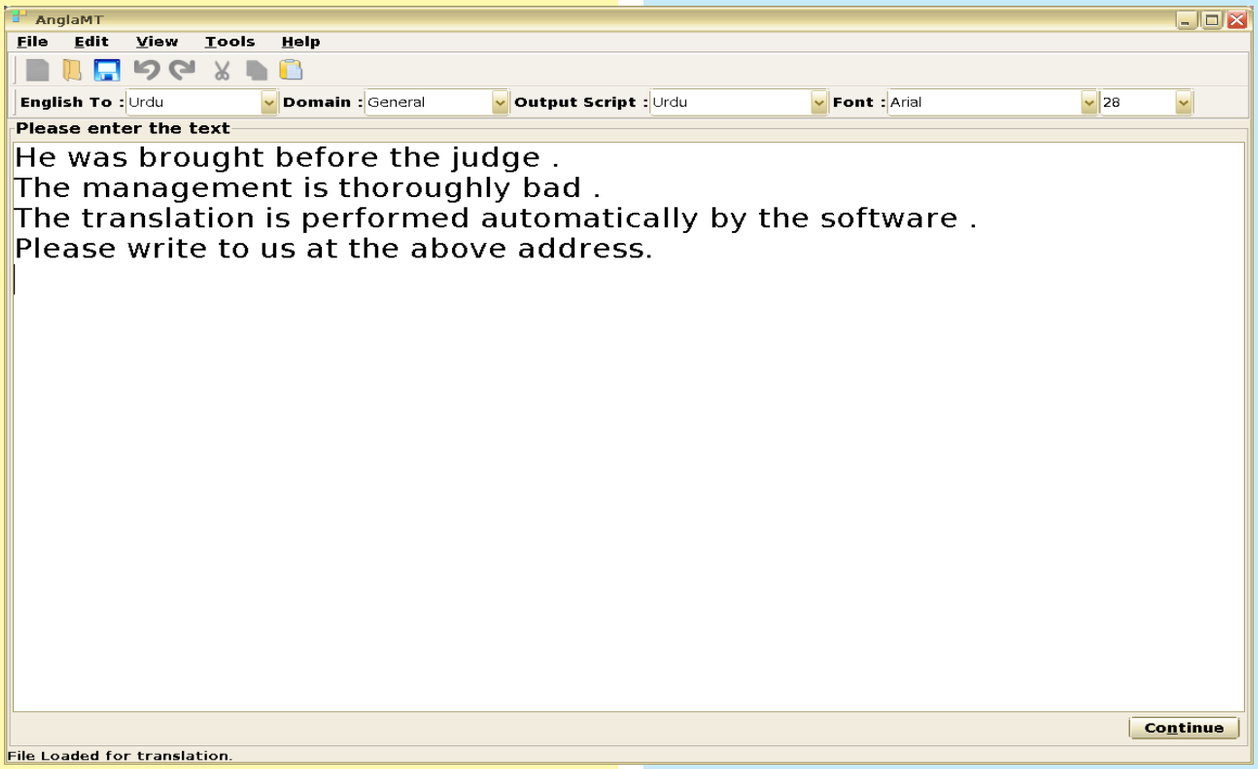


Fig.1 Input Sentence(s) Screen

Fig 2. Translation Screen

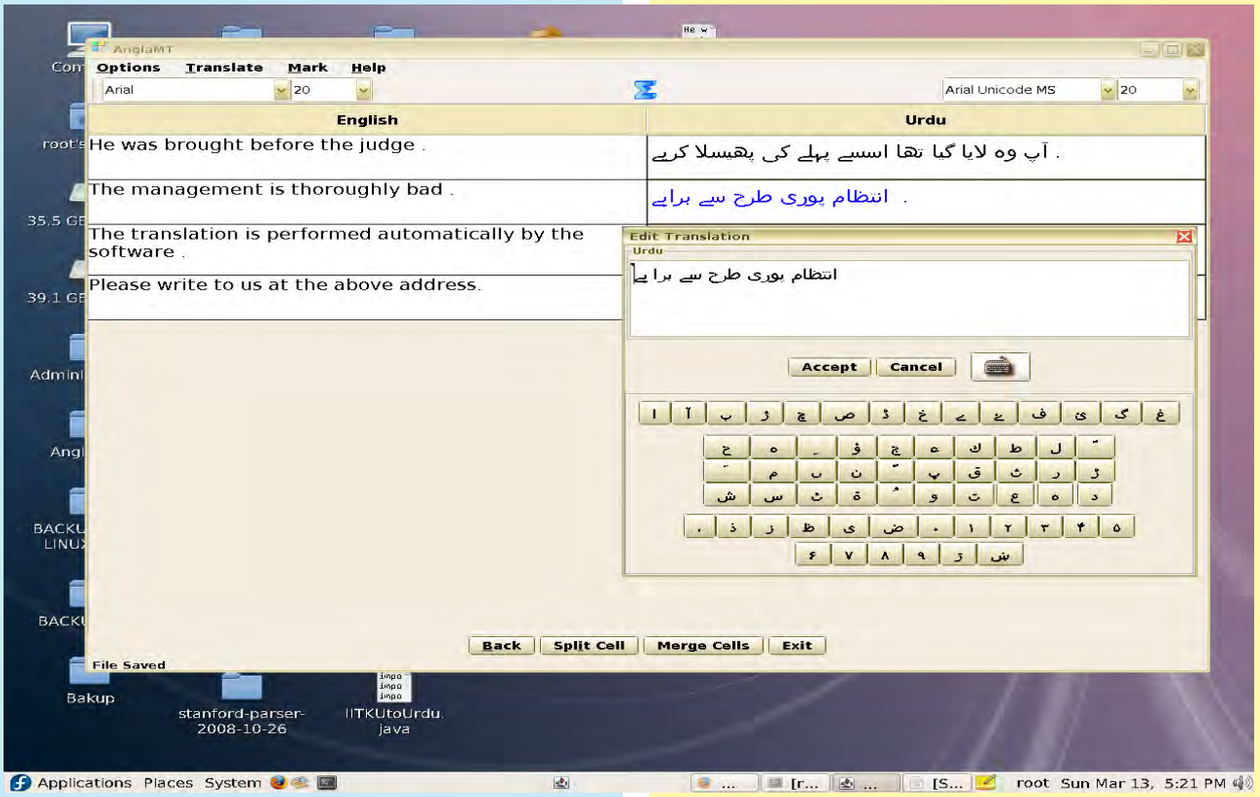


Fig. 3. Edit Translation

VII Strengths & Weaknesses of Itable approach

Strengths

1. Provides quick working translation
2. The efficiency of the system is ascertained at least up to the Hindi translation level.
3. This mechanism can be applied for the sister languages of Hindi like "Shahmukhi" language as it takes the advantage of the closeness of the two languages.
4. Less effort will be put as it may be called as a direct mapping approach.
5. Better approach for translation as it caters all the paradigm changes of Urdu as required in the translation.

Weakness

(1) The named entities obtained in Hindi are being translated during Hindi to Urdu translation, though they should have been transliterated.

S.No.	English Word (Named Entity)	Transliterated Hindi word	Translated Urdu Word
1	Jeevan	जीवन	زندگی (जिन्दगी)
2	Kavita	कविता	نظم (नज्म)

Ex. English : My name is Kavita.
Hindi Translation : मेरा नाम dfork है.
Urdu Translation ; میرا نام کویتا ہے
मेरा uTe नाम है.

This translation of named entities can be recognized for the exception to translate through Itable by marking certain NE text in the output before sending to ITable search .

(2) The searching algorithm will make the system relatively slow as it forces the system to do the extended processing or searching.

(3) For disambiguation no use of language modeling due to the lack of corpus availability.

(4) This approach only caters to or define the Hindi software as the base software. The sentences which are not parsed through the rule base or have any other problem will have the same issue in the Urdu version.

VIII Conclusion

AnglaUrdu System accepts unconstrained texts. The text may be made up of headings, texts under quotes/parenthesis, currencies, numerals, roman numbers. Proper preprocess analysis needs to be done for AnglaUrdu. Although results are satisfactory, there is a need to make the system intelligent enough to automatically learn from new translations, thereby reducing human effort from retranslating the same sentence(s)/ segment(s). The system yields good accuracy for simple sentences. More research is needed to cover complex and compound sentences with Urdu flavor and fluency in the output. The AnglaMT-I development shows the requirement to reduce the translation alternatives and reorder the best choice available to the top by using the statistics of language modeling with the target language corpus. The performance of the translation systems is a major concern. AnglaUrdu transliteration for Hindi to Urdu needs to be improved. AnglaUrdu system should be based on an Hybrid approach containing Inflectional as well as paradigm approach. AnglaUrdu system has been web enabled and is available at URL: http://tdil-dc.in for free translation.