ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

English to Kashmiri Translation System:Using Example Based Machine Translation Approach

Mutasif Ishfaq[1], Khushboo Bansal[2]

[1] M.Tech Student, Dept. of CSE, Desh Bhagat University, Mandi Gobindgarh, Punjab, India

[2]Assistant Professor, Dept. of CSE, Desh Bhagat University, Mandi Gobindgarh, Punjab, India

ABSTRACT: Machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation quality. The field of Machine Translation is responsible for the conversion of data from one natural language to other without human Intervention. This paper presents the conversion for simple English Assertive sentences to Kashmiri sentences. This is basically a machine translation. Unfortunately Kashmiri language which is a scarce resourced language has not taken into account. In this proposed system we are going through various processes such as morphological analysis, part of speech, local word grouping, for converting the meaning of simple assertive English sentences into corresponding Kashmiri sentences.

KEYWORDS*:* Artificial intelligence, Machine translation, Kashmiri Language, Morphological analysis, Empirical technique.

I. INTRODUCTION

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Kashmiri). India is rich in languages where the spoken language changes after every 50 miles. India has 22 officially recognized languages and around 2000 dialect. A number of translation systems have been developed and many systems are still developing for these. Unfortunately Kashmiri language which is a scarce resourced language has not taken into account [2][3].The Kashmiri language is one of the 22 scheduled languages of India, and is a part of the Sixth Schedule in constitution of the Jammu and Kashmir. Along with other regional languages mentioned in the Sixth schedule as well as Hindi and Urdu.

Kashmiri popularly known as Koshur is a language from the Dardic subgroup of the Indo-Aryan languages and it is spoken primarily in the Kashmir Valley, in Jammu and Kashmir. There are approximately 5,527,698 speakers throughout India, according to the Census of 2001. Most of the 105,000 speakers in Pakistan are émigrés from the Kashmir Valley after the partition of India. It is written right -to- left, top-to-bottom of page(sameas Urdu). Though the vocabularies are quite difficult at first,butto some extent there are similarities with Urduas exemplified by the following words in Table 1. They include a few speakers residing in border villages in Neelam District. Kashmiri is one of the official languages of India, and is taught in all schools in the Kashmir valley. The Kashmiri language is to be developed in the state. Most Kashmiri speakers use Urdu or English as a second language. Since November 2008, the Kashmiri language has been made a compulsory subject in all schools in the Valley up to the secondary level.

DOI: 10.15680/IJIRCCE.2015. 0310138

Language	Words				
English	Bone	Day	light	blue	new
Kashmiri	Aedij(أڑِج)	Doh (دوٚہ)	Gaash (گاش)	Nyul (نیوٗل)	Nav (نٔو)
Urdu	Haddi	Din	roshni	neela	naya

Table 1:Comparison of the similaritiesbetween different languages.

II. RELATED WORK

In the literature survey various papers has been studied to have knowledge for the NLP and the problems has been concluded that is to be solved. In the survey, that is drawn to work on the script of Kashmiri language, the main focus is on to reduce the ambiguity from the words. The ambiguity can be generated in the words where two words can have same meaning. In those kind of words there can be ambiguity because the system can-not understand the meaning of those words and can produce different results for those words. From the above literature survey it is concluded that there are many advantages and disadvantages of NLP. It helps us to choose the different ways to remove the various problems which occurs in synonyms and we use different algorithms and techniques to overcome these problems. We get knowledge about the ambiguities which occurs in Kashmiri language we apply our results on different words and then compare it with different techniques. Advantages of NLP are as follows:

Asad Abdul Malik 2013,Urdu to English Machine Translation using Bilingual Evaluation Understudy[2] We evaluated the automated machine translated outputs using Bilingual Evaluation Understudy (BLEU). The EBMT approach produced the highest accuracy of 84.21% whereas the accuracy of the online SMT system is 62.68%. We found that BLUE scores of machine translated long Urdu sentences are low in comparison with long sentences. Similarly source text containing low frequency words affect the quality of Urdu machine translation negatively.

Abbas Raza Ali, 2009,English to Urdu Transliteration System[3] This paperdescribes English to Urdu transliteration system Urdu language processing applications encounter non-Urdu text specifically English text frequently. The accuracy of these systems e.g. machine translation text-to-speech etc. is highly undermined as they are unable to handle English text. One possibility could be addition of multilingual language processing capabilities in Urdu language processing applications so that they may handle.

Abhay Adapanawar,2013, English To Marathi Translation Of Assertive Sentences [4]This paper presents the conversion for simple English Assertive sentences to Marathi sentences. This is basically a machine translation. In this proposed system we are going through various processes such as morphological analysis, part of speech, local word grouping, for converting the meaning of simple assertive English sentence into corresponding Marathi sentence.

Devinder Brar,2014,A Review of Transliteration system from English to Urdu. Term transliteration means to produce the results from source noun into target noun keeping its pronunciation same. Maximum accuracy of existing transliteration system is 63% which needs further improvement. N-Gram is used up to six-gram which has to extend to nine-gram to obtain accurate results. A web based system is required to transliterate proper nouns so that it can be used anywhere in the world.

III. PROPOSED WORK

The people from rural area are not able understand high level of English. And also it’s not possible for them to carry dictionary for conversion everywhere every time. Also these people are hesitating while using internet. Hence they can’t able to help themselves for using services like internet banking or other. Therefore, the purpose behind this project is to help people who had done their primary education but not up to the level of understanding each and every meaning of English word.

Objectives

- The objectives of this study are:
- iTo study Kashmiri and English language and their divergences.
- ii To develop a module to translate English words to their Kashmiri equivalents.
- iii To develop an English to Kashmiri Dictionary.

Challenges During English To Kashmir Translation

MT across the languages is a challenging task for several reasons like, the difference in the structure of source and target languages, ambiguity, multiword units like idioms, phrases and tense generation and many more. Some of the major challenges faced in development of English to Kashmiri MT system are as follows.

- ➢ Word ordering is different for Kashmiri and English. In Kashmiri, word order is Subject-Object-Verb (SOV) whereas in English, it is Subject-Verb-Object (SVO). Lexical differences also exist in these two languages as in some cases a group of words used in Kashmiri has a single-word equivalent in English.
- ➢ Articles are used in English but not in Kashmiri. The articles can be added at the time of post processing to correct the sentence in some cases.
- ➢ Lack of lexical resources such as digital bilingual dictionary, Tagged Corpus etc. There is no machine readable dictionary available for English to Kashmiri which can be directly used for translation, however dictionaries are available to explain the meaning of a word.
- ➢ Kashmiri is free-word order language, so it was a challenging task to identify the phrase performing the function of subject in the sentence.
- ➢ Output of the translator needs some grammatical correctness.

IV. METHODOLOGY

The methodology of a typical EBMT system for English to Kashmiri MT can be divided into four phases i.e. sentence fragmentation, search in corpus, N-ary Product based Retrieval and ordering of Translated Text as illustrated in Figure

- ➢ **Sentence Fragmentation :**Division of input sentences into phrases is vital to improve the scope of input sentences that can be handled by a translator. Same result can be achieved alternatively by keeping sentences in corpus and by gaining a broad coverage by fragmentation and combination to get new sentences using the genetic algorithm at run time. The problem of fragmenting a sentence into simpler sentences and phrases is handled using idioms, connecting words and the cutterpoints.

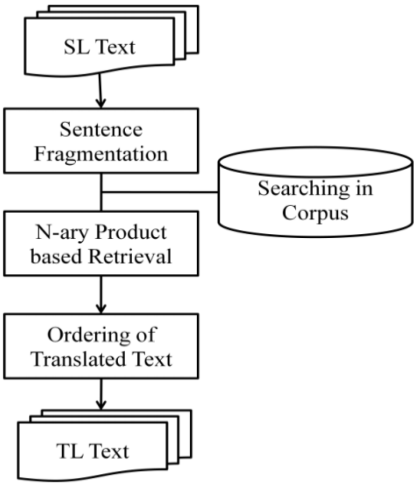


Fig 1 Flow Of Translation

- **Searching in Corpus:**In this phase the bilingual corpus is searched to determine whether the input phrase is obtainable or not. If exact match is not available, then it tries to locate the closest match. Closeness is measured via threshold at two levels i.e. for exact match and for a close match. This is achieved in two ways using Levenshtein Algorithm and Semantic Distance Algorithm.
- **N-ary Product Based Retrieval :**This phase consists of steps used to retrieve the translation of input text. For an input sentence there is a chance of getting more than one translation. The possibilities are computed and n-ary product is used to list all possible sentences.
- **Ordering of Translated Phrases:**When a single input sentence is divided into pieces and translated into output language phrases, then ordering of the translated phrases according to syntactical structure of target language is required before generating the final output. Such a process of ordering is carried out in this phase.

Algorithm:

Template matching technique is used to syllabify English transcription. In this technique syllabication is done by matching template of the form $C_{0,1}$.V.C_n[4]. Kashmiri allows only one consonant in the onset position and multiple consonants can come in the coda position of the syllable.

```
1. Convert the entire phonemic transcription of the word to consonant-vowel pairs
2. Start from the end of the word, traverse backwards to find the next vowel
3. Repeat
4. If there is a consonant preceding it?
5. Mark a syllable boundary before consonant
6. Else
7. Mark the syllable boundary before this vowel
8. End if
9. until the entire string is consumed.
```

Transformation Algorithm:

Basic difference between two languages is the structure; English is fixed order SVO while Urdu is relatively free order SOV language. If each node of VP is swapped recursively, SOV structure can be obtained. Noun phrase in both languages follows the same rule; therefore swapping is not ap plied to it. This way we obtain a basic SOV structure Exceptions of this swapping rule are present and handled. Few exceptions are: If the subject of noun phrase (NP) comprises of NP prepositional phrase (PP), transformer swap it, since the placement of PP in Urdu is before NP. If adverb phrase (ADVP) appears before verb, swapping is not needed. ADVP in English can appears in different order depending on the type of ADVP, however, Urdu prefers ADVP before verb.

DOI: 10.15680/IJIRCCE.2015. 0310138

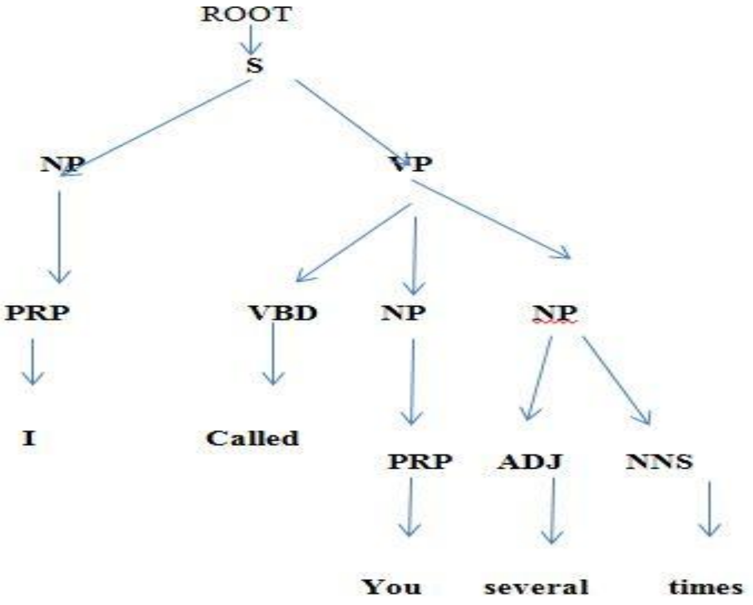


Fig 2 English Parse Tree (SVO)

Second phase is the agreement between NP and VP of the sentence S. Verb phrase in Urdu is inflected according to the gender, number and person (GNP) attribute of the head noun while form of the NP depends on the tense, aspect and modality of the verb phrase (VP), Urdu's adjectives are also modified by the GNP of head noun. Context free grammar (CFG) identifies structural attributes of language, and we need annotations in the CFG for the unification between subject and object, tree transformation algorithm defined is the generalized CFG of Urdu structure, attributes for the unification is identified, lexemes are stored in dictionary with these attributes and CFG is then annotated with rules for unification.

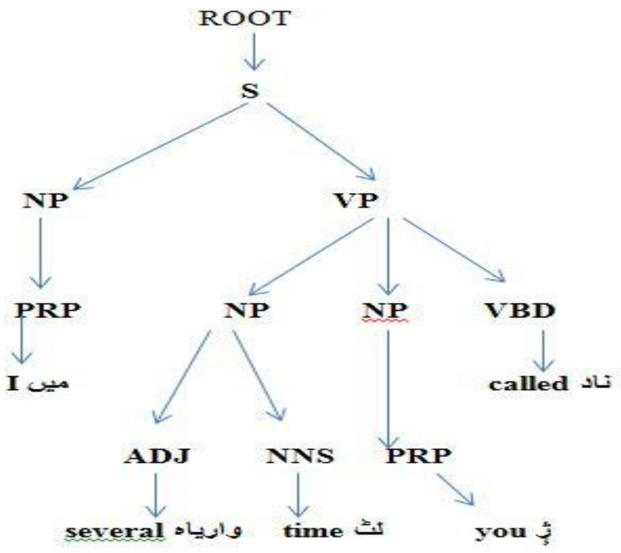


Fig 3 Kashmiri Parse Tree (SOV)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

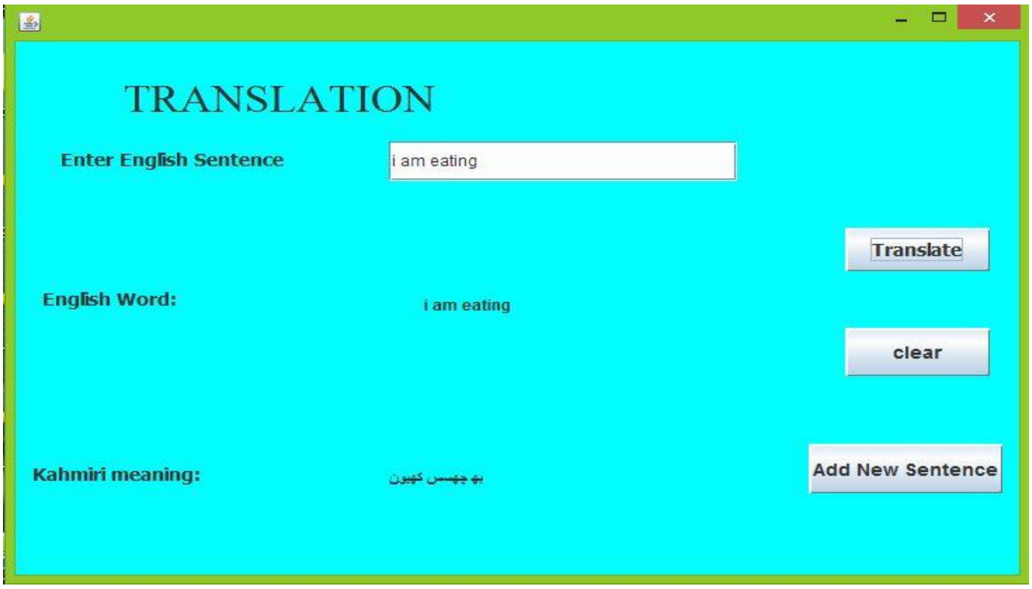
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

V. RESULT

In this section the results generation and the implementation phase in the form of GUI is represented. In the figure defined below this is a system that is representing the GUI part of the research and is defined below. It represents the GUI (Graphical User Interface) part of the research system



TRANSLATION

Enter English Sentence i am eating

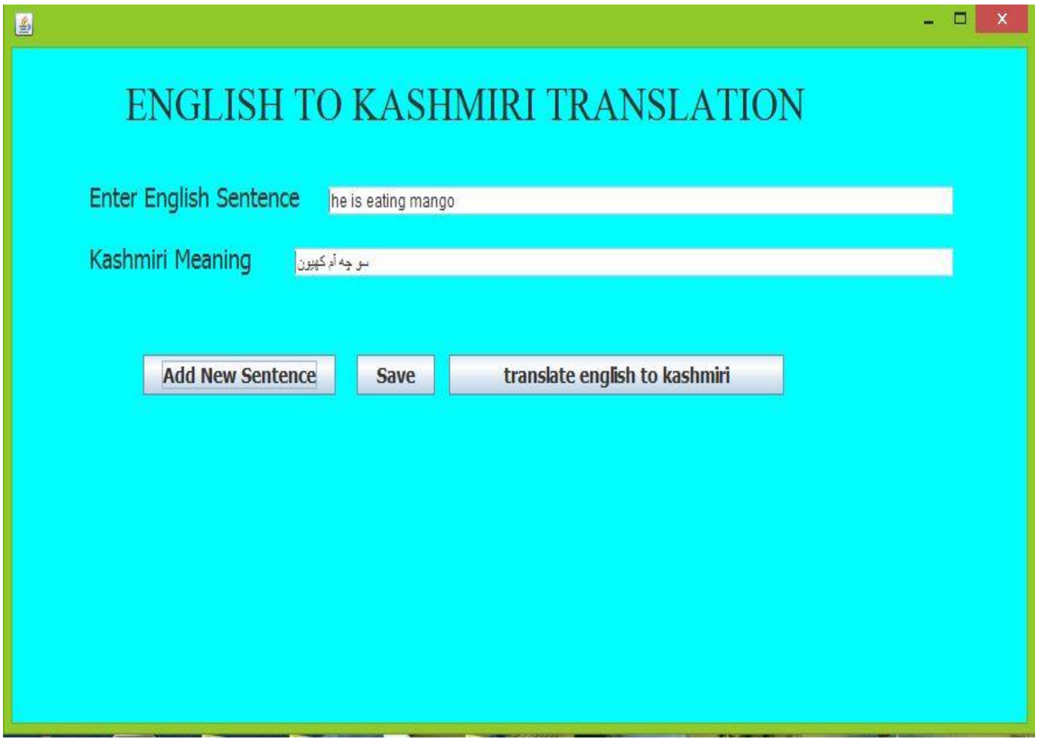
English Word: i am eating

Kahmiri meaning: بہ چھس کھیون

Fig 4 Translation from English to Kashmiri

From the above fig 4, it shows that we input English sentence and when we click translate button the English sentence gets translated into corresponding kashmiri sentence, if the input sentence is in database but, if the input sentence is not in the database then it shows a message box showing message “it doesn’t exist in database”.

From fig 5 it shows that user can also add new English sentences into database with their corresponding Kashmiri meanings.



ENGLISH TO KASHMIRI TRANSLATION

Enter English Sentence: he is eating mango

Kashmiri Meaning: سو چه آم کھیون

Fig 5 Add New Sentence to database

VI. CONCLUSION AND FUTURE WORK

Conclusion: After reviewing related papers we concluded that not much work is done for English to Kashmiri translation. Further improvements can be done in this translation system from English to Kashmiri. One of major weakness of translation from English to Kashmiri is dealing with multiple mapped characters as discussed earlier. Multiple-mapping leads to some problems in translation process. This problem also affects the accuracy of this translation system. We have developed a system which translates English sentences into Kashmiri equivalents.

Future Work: In the future scope the accuracy of the system in the script of Kashmiri language can be improved. As no work has done in translation from English language to Kashmiri language So, Prediction ability of the research system can be improved. In the prediction ability of the system the System will be able to produce the accurate results from English to Kashmiri language. In future, quality can be improved to increase the size of corpus. One more future work to be done for this System is to make it automatic to translate from English to Kashmiri and vice-versa.In future one more thing to be done with this system is to make it an online dictionary for English to Kashmiri language.

REFERENCES

[1] http://en.wikipedia.org/wiki/Kashmiri_language

[2 Asad Abdul Malik,Asad Habib. "Urdu to English Machine Translation using Bilingual Evaluation Understudy ".International Journal of Computer Applications (0975 – 8887) Volume 82 – No 7, November 2013.

[3] Abbas Raza Ali,MadihaIjaz."English to Urdu Transliteration System ".Proceedings of the Conference on Language & Technology 2009.

[4] AbhayAdapanawar,Anita Garje,Paurnima Thakare,PrajaktaGundawar,Priyanka Kulkarni."English To Marathi Translation Of Assertive Sentences".International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 2, March -April 2013, pp.516-518.

[5] Devinder Brar, Er. Rishamjot Kaur."A Review of Transliteration system from English to Punjabi". Volume 4, Issue 7, July 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.

[6] Vaishali Gupta,Nisheeth Joshi,ItiMathur."Subjective and Objective Evaluation of English to Urdu Machine Translation".In Proceedings of the Third Workshop on Statistical Machine Translation, pages 115–118. June 2008.

[7] Saleem, M. “Urdu RasmulkhatkiJaamiat”. cases, the custom built fonts have to download prior to viewing the page. Sending an Kashmiri e-mail through ordinary mail accounts is not possible Akhbar-i-Urdu, Pages 6-10, Islamabad, Pakistan, 2002.

[8] R. Bokhari, and S. Pervez, “Syllabification and Re-Syllabification in Urdu”.Akhbar-i-Urdu, Pages 63-67, Islamabad, Pakistan, 2003.

[9]) Mr. Krushna Belerao, Prof. V.S.Wadne “Machine Translation Using Open NLP and Rules Based System “English to Marathi Translator”” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 6

[10] AbhijeetR.Joshi, M. Sasikumar, “Constructive approach to teach inflections in Marathi language”,www.cdacmumbai.in/design/corporate_site/.../pdf.../CATIML1.pdf

BIOGRAPHY

The author name is MUTASIF ISHFAQ.He has done B.Tech in Information Technology Engineering from Baba Ghulam Shah BadshahUniversity,Rajouri,J&K with first grade. Currently he is pursuing M.Tech in Computer Science Engineering from Desh Bhagat University. Mandi Gobindgarh, Punjab.The author has worked on various platform such as ASP.NET,JAVA,PHP.His research intrests are Natural language Processing,networking,Crytography.