Resource Development for English to Gujarati Machine Translation system

RESOURCES DEVELOPMENT FOR ENGLISH TO GUJARATI MACINE TRANSLATION SYSTEM

C. K. Bhensdadia, Brijesh Bhatt Jatayu Baxi,
Kirit Patel and Dinesh Chauhan Department of Computer Engineering,
Faculty of Technology Dharmsinh Desai University, Nadiad. Gujarat.

Abstract

This article describes various linguistic resources created to develop English to Gujarati Machine Translation system. The work includes, parallel corpus creation, English-Gujarati Lexicon building and development of grammatical resources such as Transfer Grammar and Morph Synthesizer. The resources are used to develop Tree Adjoining Grammar based English to Gujarati Machine Translation system. The system shows 50 % accuracy on gold data.

1. Introduction

Machine Translation refers to build software system which translates text from one natural language to another natural language.1 In a multilingual nation like India with 22 official languages, it is important to translate and share information across languages. Most of the information available online is in English. In order to make this information available to common Citizen of India, it is desirable to translate this information in vernacular languages. With this aim, English to Indian Language Machine Translation System (EILMT) project is initiated by Department of Information Technology, MCIT, Government of India. Aim of EILMT project is to design and deploy a Machine Translation System from English to Indian Languages. The project started from September 2006. Figure 2 shows abstract view of E-ILMT system. Work for Gujarati language is initiated in 2011. Remaining of the paper is organized as follows, section 2 describes basic over view of Machine Translation system and various approaches For the development of Machine Translation system. Section 3 provides brief description of EILMT system and interface for Gujarati language. Section 4 and 5 describe development of resources for Gujarati language.

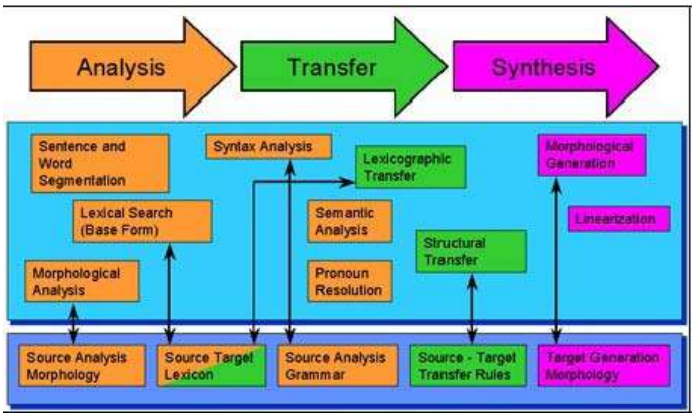

Figure 1: Basic Architecture of Machine Translation (Image source: http://www.linguatec.net/products/tr/information/technology/mtranslation)

2. Overview of machine translation

2.1 Various Approaches of Machine Translation

'A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called Grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules has discussed knowledge based machine translation in which system rely on set of language pair-dependent rules to carry out translation. Shortcoming of this approach is insufficient amount of really good dictionaries.

2.1.1 Rule-based Approach

In this section we discuss various approaches to build Machine translation system describes survey done on various approaches for machine translation system.

2.1.2 Statistical Machine Translation

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora proposed an effective frame-work for English-Hindi phrase-based SMT. In this work With only a small amount of bilingual training data and limited tools for Hindi, reasonable performance and substantial improvements over the base-line phrase-based system is achieved. Shortcoming of this approach is Corpus creation can be costly for users with limited resources.

2.1. 3 Example Based Approach

'An EBMT system is given a set of sentences in the Source Language and their corresponding Translations in the Destination Language, and uses those examples to translate other, similar sentences Makoto Nagao proposed this method and pointed out that the Example-based Machine Translation is especially adapted to the translation between two totally different languages.

2.1.4 TAG Based Approach

Tree-adjoining grammar (TAG) is a grammar formalism defined by Aravind Joshi. Tree-adjoining grammars are similar to context-free grammars, but the elementary unit of rewriting is tree rather than the symbol. Whereas context-free grammars have rules for rewriting symbols as strings of other symbols, tree-adjoining grammars have rules for rewriting the nodes of trees as other trees. A TAG based system system has three phases i) Analysis ii) Transfer and iii) Generation. In the first stage source language parser generates syntactic representation of a sentence. In next stage result of first stage is converted into target language oriented representations. In the final step of this translation approach, a Target Language morphological analyzer is used to generate the final Target Language output.

3. E-ILMT System

Figure 2 shows basic architecture of E-ILMT system. Input is first pre processed with help of modules like Morph analyzer, Named entity recognizer, word sense disambiguation etc. After that there are three approaches which are tried in this system: Example based machine translation, Statistical machine translation and tag based translation. For EBMT model, training examples are provided to the system, for SMT appropriate language model is prepared. For tag based approach parser and generator module is implemented. Post processing modules like Morph synthesizer, multiple output selector, Synonym

selector are also implemented. The ranking module is also implemented which gives rank to each output generated with different approaches.

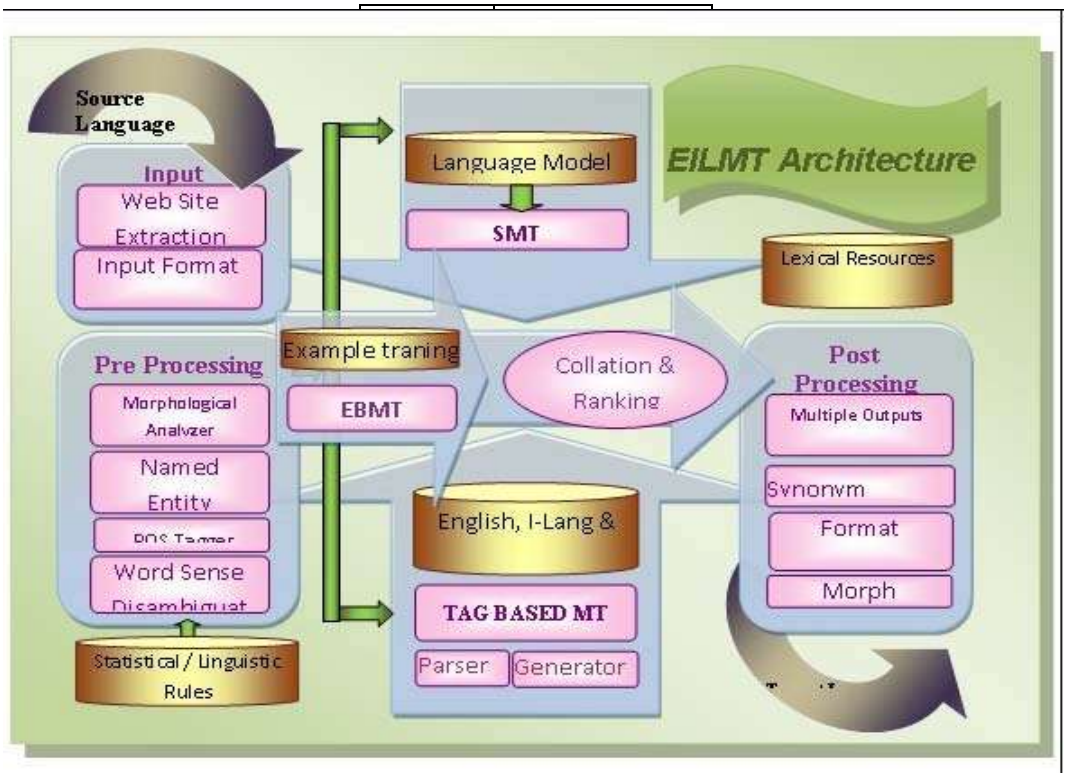

Figure 2: Basic Architecture of E-ILMT System

6. Corpus Development for Gujarati

In order to integrate Gujarati language support in E-ILMT framework various resources and tools are developed for Gujarati language. The existing tourism and health corpus is translated into Gujarati to create parallel corpus. The corpus is XMLized and a bilingual English-Gujarati lexicon is created. The morphological structure of Gujarati is investigated and accordingly morph synthesizer for Gujarati is developed. This section describes various tools and resources created for Gujarati language.

4.1 Translation

Domain Domain	No of Sentences
Health	15000
Tourism	12000

Table 1: Corpus Creation

The tourism and health corpus which were available in other languages were used for the system development. The first step was to translate English tourism and health corpus in Guajarati language to construct parallel corpus. The sentences are categorized based on their structure into Simple, Complex, Copula, Adverb-Initial, Gerund etc.

As shown in Table 1, we translated 15000 sentences of health domain and 12000 sentences of tourism domain. The sentences are grouped into files each containing 100 sentences. With each English sentence we put corresponding Guajarati translation of the sentence and put it in the same file. Figure 3 shows sample parallel corpus file.

S.N.	English Sentences	Gujarati Sentence
1	A hot Epsom salt bath, twice a week will be highly beneficial in all cases of acne.	અઠવાડિયામાં બે વખત સિંધવમીઠાવાળા ગરમ પાણીનું સેવન ખીલના બધા કિસ્સાઓમાં લાભદાયી રહેશે.
2	A person can, through chronic stress, become sensitive to common foods or commonplace substances like petrol fumes.	લાંબા સમયથી ચાલતા તણાવદ્વારા એક વ્યક્તિ સામાન્ય ખોરાક અથવા પેટ્રોલ જેવા સામાન્ય જગ્યાઓ પર રહેલ પદાર્થો પ્રત્યે સંવેદનશીલ બની શકે છે.

Figure 3: Parallel Corpus Creation

4.2 XMlization

E-ILMT project uses XML as a standard language to represent parallel corpus. In order to represent Guajarati corpus into XML, we used XMLization tool developed at Banasthali Vidyapith. The XMLization tool takes parallel corpus file as an input and generates two XML files, one for source (English) language and other for target (Gujarati) language. Figure 4 shows an example XML file.

```
<segment segmentid="1">

<sentence sentencenumber="1" subdomain="HL" >
A hot Epsom salt bath, twice a week will be highly beneficial in all cases of acne.
</sentence>
<sentence sentencenumber="2" subdomain="HL" >
A person can, through chronic stress, become sensitive to common foods or commonplace substances like petrol
fumes.
</sentence>
<sentence sentencenumber="3" subdomain="HL" >
A person generally takes to drinking as a means to enliven social life, to overcome anxiety or to induce sleep.
</sentence>
<sentence sentencenumber="4" subdomain="HL" >
About 90 percent of the alcohol is slowly oxidised in the liver and the remaining 10 percent is eliminated by
breathing and through urination.
</sentence>
```

Figure 4: XML file Creation

4.3 Lexicon Building

The Lexical transfer phase of machine translation finds a target language word for the given source language word. English to Gujarati lexicon is constructed to perform this task. The Linguistic Resource Management Tool (LRMT tool) developed by IIIT-Allahabad is used to construct lexicon.

The LRMT tool opens XML file and displays English and Guajarati sentences parallaly. The words which are not present in dictionary are highlighted. We need to select source word and corresponding destination language word and click Add word button. After we add word, it will be added into database. Figure 5 shows a snapshot of the LRMT tool. Also we can add features with Noun and add synonyms of given noun. Following features are identified and added with each Noun.

- Proper Noun : Indicates weather Noun is proper noun or simple Noun.
- Gender : Indicates gender of Noun. Gujarati has three genders so it can take one of GM,GF or GN value.
- Number : Indicates weather Noun is singular or Plural.
- Person : Indicates person of Noun. Either PI,PII or PIII.
- Animate : Indicates weather Noun is Animate or Inanimate.
  Human : Indicates nature of noun as Human or Non- Human.
- Abstract : Indicates Abstract or Non-Abstract Noun.
- Honorific : Indicates Honorific nature of Noun.
- Temporal : Indicates Temporal or locative nature of Noun.
- Countable : Indicates weather Noun is countable or not.

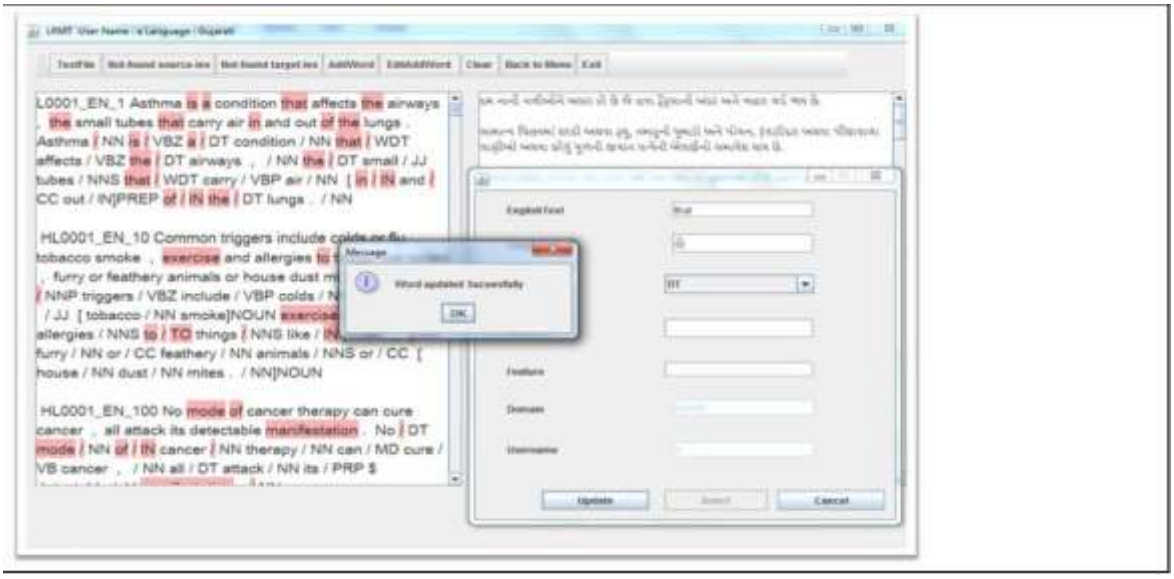

Figure 5: Lexicon Creation using LRMT Tool We have done lexicon building and feature addition for around 3200 words

5. Tool Developement

5.1 Transfer grammer

Transfer grammar module chooses appropriate Gujarati word for the given English verb. English verbs are categorized into 6 parts. Transfer Grammar rules are developed to category wise relates English and Guajarati verbs. In Gujarati, verb shows gender inflection, which is not the case with English. As shown in Table 2, the noun shows no inflection for gender but in Table3, verb shows inflection ો for gender male and ી for gender female. Transfer grammar reflects these inflection accurately in translation. Figure 6 shows an example of transfer grammar entry for verb type Appoint. We have built around 900 rules for transfer grammar.

Verb Type	English Sentence	Gujarati Translation
TYPE APPOINT	A boy did the work.	છોકરાએ કામ કયુળઋ.
TYPE APPOINT	A Girl did the work.	છોકરીએ કામ કયુળઋ.

Table 2: Verb type Appoint

Verb Type	English Sentence	Gujarati Translation
TYPE ALLOW	A boy came.	છોકરો આષઋયો.
TYPE ALLOW	A Girl came.	છોકરી આવી.

Table 3: Verb type Allow Example

કરવા-પડશે^#^0,GM,NP,PIII,0,0,0,-H,0,0!!!કરવો-પડશે^#^0,GM,NS,PIII,0,0,0,+H,0,0!!!કરવી-પડશે^#^0,GF,NS,PIII,0,0,0,-H,0,0!!!કરવી-પડશે^#^0,GF,NP,PIII,0,0,0,-H,0,0!!!કરવી-પડશે^#^0,GF,NS,PIII,0,0,0,+H,0,0!!!કરવું-પડશે^#^0,GN,NS,PIII,0,0,0,+H,0,0!!!કરવા-પડશે^#^0,GN,NP,PIII,0,0,0,+H,0,0	કરવા-પડશે-નહિ^#^0,GM,NP,PIII,0,0,0,-H,0,0!!!કરવો-પડશે-નહિ^#^0,GM,NS,PIII,0,0,0,+H,0,0!!!કરવી-પડશે-નહિ^#^0,GF,NS,PIII,0,0,0,-H,0,0!!!કરવી-પડશે-નહિ^#^0,GF,NP,PIII,0,0,0,-H,0,0!!!કરવી-પડશે-નહિ^#^0,GF,NS,PIII,0,0,0,+H,0,0!!!કરવું-પડશે-નહિ^#^0,GN,NS,PIII,0,0,0,+H,0,0!!!કરવા-પડશે-નહિ^#^0,GN,NP,PIII,0,0,0,+H,0,0

Figure 6: Sample Transfer Table Entry

5.2 Morph Synthesizer

A morph synthesizer is a tool which synthesizes output word according to its grammatical features. As destination Gujarati language is inflectional, we need to identify various inflections and also build the rules regarding when to apply which kind of inflection. In this system the root word is stored in database and while translating, based on grammatical features of source word, destination word is inflected and hence synthesized output is produced.In this section we desribe rules for each of above mentioned synthesizer and implementation details for the same.

For this project we have built rules and implemented those rules for following synthesizers :

5.2.1 OF Synthesizer

The task of OF synthesizer is to replace corresponding target language inflected word for “OF”. In this synthesizer, English sentence having following structure is scanned:

NN1+OF+NN2

NN1 and NN2 are Nouns. These nouns have some features associated with them. Based on feature of NN1, the rule is applied on NN2 and output word is synthesized.

For example if English sentence is: Book of Ram. So here NN1 is Book, NN2 is Ram. Feature of NN1 is GF(Gender Female) , so according to table 4 we apply the rule and append ની to NN2 and hence translation turns out to be રામની ચોપડી.

Feature of NN1	Rule applied on NN2
GM	NN2+નો
GF	NN2+ની
GN	NN2+નું
NP	NN2+ના

Table 4: Rules For OF Synthesizer

5.2.2 Adjective Synthesizer

The task of Adjective synthesizer is to synthesize adjective in target language based on the noun that an adjective follows. The English sentence is in following format :

ADJ+NN

Table 5 shows the rules applied on Adjective based on feature of NN.As an example if the sentence is Good boy then ADJ is good and NN is Boy. So as Boy is having feature GM(Gender Male) , the corresponding rule is applied and output is સારો છોકરો.

5.2.3 Apostrophe Synthesizer

The purpose of Apostrophe synthesizer is to synthesize output for the sentences which includes Apostrophe s. The format of such sentences are :

Feature of NN	Rule applied on ADJ
GM	(ADJ-2)+ો
GF	(ADJ-2)+ી
GN	ADJ
NP	(ADJ-2)+ા

Table 5: Rules For Adjective Synthesizer

NN1+Apostrophe s+NN2

Table 6 shows the rules applied on NN1 based on features of NN2.For example English sentence is Ram's Book. So NN1 is Ram and NN2 is Book. As features of NN2 is GF so Rule 2 is applied and output is given as રામની ચોપડી.

Feature of NN2	Rule applied on NN1
GM	NN1+નો
GF	NN1+ની
GN	NN1+નું
NP	NN1+ના

Table 6: Rules For Apostrophe Synthesizer

6. Observation

After above synthesizers are implemented in the system, we have prepared 100 Gold sentences which contains different types of sentences and tested the system for Gujarati. We kept manual translation as reference and evaluated accuracy of output on the scale of 0 to 5.The average rating of output out of 5 is around 2.5. So we can conclude that accuracy of the system is approximately 50%.

7. Conclusion

Gujarati language is successfully included in EILMT system. The system is tested for TAG based approach. It shows around 50% accuracy on the gold data. We aim to further improve the performance of the system by increasing lexicon size and investigating TAG structure for Gujarati. We also aim to build statistical Machine Translation system for English to Gujarati machine translation.

8. Acknowledgements

English to Gujarati Machine Translation system is developed as a part of 'Anuvadaksha' project. The support of Ministry of Communication and Information Technology, Government of India is gratefully acknowledged.

References

1. Antony P. J.,Machine Translation Approaches and Survey for Indian Languages,Computational Linguistics and Chinese Language Processing,Vol. 18, No. 1, March 2013, pp. 47-78.

2. R. Ananthakrishnan, Jayprasad Hegde, Pushpak Bhattacharyya, Ritesh Shah and M. Sasikumar,Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation, International Joint Conference on NLP (IJCNLP08), Hyderabad, India, Jan, 2008.

3. Makoto Nagao (1984), A framework of a mechanical translation between Japanese and English by analogy principle , In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers.

4. Nirenburg, Sergei (1989), Knowledge-Based Machine Translation, Machine Trandation 4 (1989), 5 -24. Kluwer Academic Publishers. Retrieved 20 June 2012.

5. Joshi Aravind, S. R. Kosaraju, H. Yamada(1969), String Adjunct Grammars, Proceedings Tenth Annual Symposium on Automata Theory, Waterloo, Canada.

✡✡✡