INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

English-Hindi Translation System with Scarce Resources

Nitin Wagadia
Bachelor of Engineering (B.E), K. J. Somaiya College of Engineering, Mumbai University, Mumbai, India
Prakash Ravaria
Bachelor of Engineering (B.E), K. J. Somaiya College of Engineering, Mumbai University, Mumbai, India
Rahul Bhat
Bachelor of Engineering (B.E), K. J. Somaiya College of Engineering, Mumbai University, Mumbai, India
Shail Parekh
Bachelor of Engineering (B.E), K. J. Somaiya College of Engineering, Mumbai University, Mumbai, India

Abstract:
Language forms the basis of human communication. There are many different languages spoken in this world among which English is the global language .Machine Translation is the translation of one natural language into another using automated and computerized means. Translation systems are very handy for the tourist. The translation becomes an easy job when the domain of translation is known. In this paper, we discuss the various approaches taken for building the machine translation system and then discuss English-Hindi Machine Translation System with scarce resources for tourist.

1. Introduction

The technology is reaching new heights, right from conception of ideas up to the practical implementation. It is important, that equal emphasis is put to remove the language divide which causes communication gap among different sections of societies. Natural Language Processing (NLP) is the field that strives to fill this gap. Machine Translation (MT) mainly deals with transformation of one language to another. Coming to the MT scenarios in India, it has enormous scope due to many regional languages of India. It is pertinent that majority of the population in India are fluent in regional languages such as Hindi, Punjabi *etc*. Given such a scenario, MT can be used to provide an interface of regional language. In this paper, we discuss the various approaches taken for building the machine translation system and then discuss English-Hindi Machine Translation System with Scarce resources.

2. Approaches

Machine translation helps people from different places to understand an unknown language without the aid of a human translator. The Source Language (SL) is the language which is to be translated & the Target Language (TL) is language in which it is translated. While translating, the syntactic structure and semantics structure of both source language and target language should be considered. The major machine translation techniques are

- Statistical Machine Translation Technique (SMT)
- Example-based machine translation (EBMT)
- Direct machine translation
- Rule Based Machine Translation Technique

2.1. Statistical Machine Translation

Statistical Machine Translation is one of the most widely used machine translation approaches in the modern era .The statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The SMT is a corpus based approach, where a massive parallel corpus is required for training the SMT systems. The SMT systems are built based on: language model and translation model. Its advantage is that linguistic knowledge is not required for building them. For machine translation of English to Indian languages, the parallel corpora have to be pre-processed (changing word-order) and trained in SMT.

2.2. Example Based Machine Translation:

The example based machine translation (EBMT) is the corpus based approach without any statistical models. These systems are trained with the parallel corpus of example sentences, similar to SMT systems. The example based systems generally don't learn from the corpus. They store the parallel corpus and uses matching algorithms to search and retrieve the sentences.

EBMT entails three steps:

- Matching fragments against the parallel corpus
- Adapting the matched fragments to the target language
- Recombining these translated fragments appropriately.

2.3. Direct Machine Translation

Direct Machine Translation is the one of the simplest machine translation approach. In Direct Machine Translation, a direct word by word translation of the input source is carried out with the help of a bilingual dictionary and after which some syntactical rearrangement are made. In Direct Machine Translation a language called the source language is given as input and the output is known as the target language. Typically, the approach is unidirectional and only takes one language pair into consideration at a time.

2.4. Rule Based Machine Translation

The Rule Based Machine Translation System takes into account semantic, morphological and syntactic information from a bilingual dictionary and grammar. These rules generate the output target language from the input source language. The rule based machine translation system is developed by hand coded rules for translation. The system requires good linguistic knowledge to write the rules and a bilingual dictionary is also needed. The rule based systems are highly suited for translation of English to Indian Languages because the bilingual dictionary could be collected easily compared to parallel corpus and the rules could also be written well with the help of linguists.

3. Proposed System

In this section the method to develop English-Hindi Machine Translation System with scarce resources is discussed. While developing Machine Translation system lots of resources is required to get good translation. These resources in turn increase the cost of the system. We have used the Rule Based approach to develop the English-Hindi MT system. The Statistical Machine Translation Systems even tough doesn't require any grammar rules or knowledge; it requires lot of resources in terms of corpus. Example based Translation system also requires huge parallel corpus to generate phrase pair. Direct translation requires minimum resources but the quality of translation is not good. The reason for using Rule Based approach is that it requires very less resources and gives comparatively better translation.

Rules Based System basically requires only three main things:

- Bilingual Dictionary
- English sentence Structure
- Rules to reorder English sentence structure to Hindi language structure.

A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrased from one language to another. We have used '*shabdkosh*' English-Hindi Bilingual dictionary to get meaning of English words.

Figure 1: System Architecture

4. Parsing

Parser is an algorithm which produces a syntactic structure for a given input. The parser is the first component of the rule based machine translation system and it is used on the source (English) side. The Parser is used for four main purposes in the machine translation system. The parser is used for syntactic analysis of the English sentence in order to give the parse tree structure of the English sentence by context free grammar. The example of parsing is shown in fig. below:

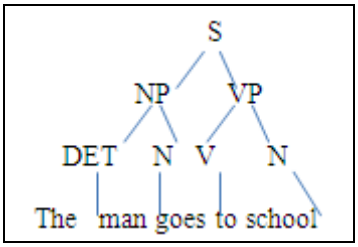

Figure 2

This tree structure is required for re-ordering the source (English) sentence with respect to the target (Hindi) sentence by transfer rules. The parser is used for Parts of Speech (POS) tagging of the English sentence to give English words and their corresponding POS tags. These POS tagged words are used to search the target equivalent of English word in bilingual dictionary. We have used Stanford Parser to get the syntactic structure of the English Sentences.

4.1. Reordering Rules

All the languages have their own structure. English Language follows Subject Verb Object (SVO) Structure whereas Hindi Language follows Subject Object Verb (SOV) structure. Hence, in order to translate English language structure to Hindi language structure we have used reordering rules [paper name]. Stanford Parser generates various phrase tags, out of which five candidates are found to be very useful for reordering. These include VPs (verb phrases), NPs (noun phrases), ADJPs (adjective phrase), PPs (preposition phrase) and ADVPs (adverb phrase).

Tag	Description(Penn tags)
dcP	Any, parser generated phrase
pp	Prepositional Phrase(PP)
whP	WH Phrase(WHNP, WHADVP, WHADJP, WHPP)
vp	Verb Phrase(VP)
sbar	Subordinate clause(SBAR)
np	Noun phrase(NP)
vpw	Verb words(VBN, VBP, VB,VPW,MD, VBZ, VBD)
prep	Preposition words(IN,TO,VBN,VPW)
adv	Adverbial words(RB, RBR, RBS)
adj	Adjunct word(JJ,JJR,JJS)
advP	Adverb phrase(ADVP)
punct	Punctuation(,)
adjP	Adjective phrase(ADJP)
Tag?	Zero or one occurrence of Tag

Table 1

The format for writing the rules is as follows:
Type_of_phrase (tag1 tag2 tag 3: tag2 tag1 tag3)
This means that “tag1 tag2 tag3”, structure has been transformed to “tag2 tag1 tag3” for the type_of_phrase. In this paper type_of_phrase denotes our category (S, SBAR) in which rule fall. The table given above explains about various tags and corresponding Penn tags used in writing these rules.

SBAR Rules:
Rule 1: SBARQ (whP SQ (VPW NP)): NP whP VPW
Sentence: How are you?
Parsing: (SBARQ (whP (WRB how)) (SQ (VPW are) (NP (PRP you)))))
Reordering: NP (PRP you) whP (WRB how) VPW (are)
Translation: AAP KAISE HO

Rule 2: SBARQ (whP SQ (VPW NP VP)): NP whP VP VPW
Sentence: What are you doing
Parsing: (SBARQ (whP (WP what)) (SQ (VPW are) (NP (PRP you)) (VP (VPW doing)))))
Reordering: NP (PRP you) (whP (WP what)) VP ((VPW doing))(VPW are)
Translation: AAP KYA KAR RAHE HAI

S Rules
Rule 1: S (NP VP (VPW1(prep VPW2)):NP VP (VPW2 VPW1prep)
Sentence: I want to eat
Parsing: (S (NP (PRP I)) (VP (VPW want) (S (VP (prep to) (VP (VPW eat)))))))
Reordering: NP ((PRP I)) VP (VPW (eat) (VPW want) (prep to))
Translation: MUJHE KHANA CHAIYE

Rule 2: S (NP1 VP(VPW1 S(prep VP(VPW2) PP(prep) NP2)) : NP1NP2 prep VPW2 PP(prep) VPW1
Sentence: I want to go to restaurant
Parsing: (S (NP (PRP I)) (VP (VPW want) (S (VP (prep to) (VP (VPW go) (PP (prep to) (NP (NN restaurant)))))))))
Reordering: NP((PRP I)) (NP (NN RESTAURANT)) (prep to)(VPW go) (PP (prep to))(VPW WANT))
Translation: MUJHE RESTAURANT JANA HAI

5. Conclusion

Given the scenario where the resources are scare and the domain for translation is known. Better results can be obtained by using Rule based approach when compared with other Machine translations System. Rule Based MT system requires knowing the structure of the input sentence. Once the structure and its transfer rules are identified the translation can be easily done with better efficiency. We have devised four rules on the basis on most commonly used sentences by tourist and have found better results.

6. References

1. Ananthakrishnan R, Kavitha M, JayPrasad J Hegde, Chandra Shekar, Ritesh Shah, Sawani Bade, Sasikumar M (2006), MaTra: a Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation, In proceedings of Symposium on Modeling and Shallow Parsing of Indian Languages.
2. Anitha T Nair, Sumam Mary Idicula (2012), Syntactic Based Machine Translation from English to Malayalam,Data Science & Engineering (ICDSE), 2012 International Conference on 18-20 July 2012, Pg. 198-202.
3. Pushpak Bhattacharyya (2012), Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality, CSI Journal of Computing, Vol 1- No 2, pg. 1-13.
4. ShwetaDubey, TarunDharDiwan(2012), Supporting Large English-Hindi Parallel Corpus using Word Alignment, International Journal of Computer Application, Vol-49 (6), pg. 16-19.
5. Raj Nath Patel, Rohit Gupta, Prakash B. Pimpale and Sasikumar (2013), Reordering rules for English-Hindi SMT, In Proceedings of the Second Workshop on Hybrid Approaches to Translation, Pg.34-41.
6. SugataSanyal, RajdeepBoroghain (2013), Machine Translation Systems in India, Computer Research Repository