

STATISTICAL MACHINE TRANSLATION FROM ENGLISH TO HINDI

¹Subalalitha, ²Aarthi Venkataraman, ³BasimShahidBaqui¹Department of Computer Science and Engineering,

SRM Institute of Science and Technology, Kattankulathur, India.

²Tata Consultancy Services²Digital Zone, No. 79, Rajiv Gandhi IT Expressway, Chennai.³Citius Software Private Limited, 7 Habibullah Road, Chennai¹subalalitha@gmail.com, ²aarthi.venkataraman@gmail.com³Basimshahid@gmail.com

Abstract: Machine Translation (MT) is one of the most popular applications of Natural Language Processing which automatically translates a source language to the target language. Especially in a country like India that has around 150 languages, it is difficult to ensure a standardized language that is understandable by all. Furthermore, with the increasing demand for knowledge on World Wide Web(WWW), there is a strong need for an MT system. This paper focuses on building a MT system, which uses a statistical approach for translating text from English to Hindi. The word orders of the target language are defined by N-gram Model and Naïve Bayes algorithm is used to predict the next word/words having the highest probability. This system has been tested and compared with an existing rule-based system and it has a recall percentage of 73.43% compared to that of the rule-based system which has 35.88%.

Key words: Machine Translation, Statistical Machine Translation, Indian Machine Translation

1. Introduction

With the enormous increase of English language pages on the World Wide Web (WWW), cross lingual access of documents become quintessential. A Machine Translation (MT) system can aid greatly in the cross lingual access, wherein, a user can read the English documents in his/her local language[1]. This paper proposes a statistical MT system, which automatically translates the English text to Hindi. Handling Indian languages and English in an MT system is quite difficult as the word orders of Indian languages are incompatible with that of the English [2]. The proposed approach has attempted to bridge that gap and has handled both simple and complex sentences.

Initially, the MT systems were based on a concept called direct translation which followed a word to word translation from the source language to the target language [3]. However, due to the number of different

languages and their varying properties and sentence structures, direct translation did not give satisfactory results. Later, came in the concept of rule-based translation, where the source language is converted into an intermediate representation, which then can be converted into the required target language. The main idea of this method is to use a statistical-based analysis to convert the source text/language into the target text/language. The major advantage of this method is that it does not depend on a dictionary or any hard coded rules. It is purely based on the calculated value of the probability, which in turn depends on the size of the training set.

A considerable amount of research has been done on MT systems involving Indian languages [4]. This has led to the development of many Indian language MT systems like:

- Anglahbharati (1995)
- Shakti (2000)
- MaTra
- Mantra
- Anusaarak
- Google Translate

It is important to note that constructing a perfect machine translation system is by far still a far-fetched dream. There does not exist a MT system that translates from one language to another with 100% perfection.

The rest of the paper is organized as follows. Section 2 describes the different approaches to MT (background). Section 3 describes the literature survey on the existing MT systems. Section 4 describes the proposed approach in detail. Section 5 explains the evaluation of the proposed approach and section 6 gives the conclusions and future work.

2. Background

There are different approaches to MT as follows and each method has various types in turn [5].

- Rule Based Machine Translation Approach(RBMT)
 - Direct Machine Translation
 - Interlingua Machine Translation
 - Transfer-Based Machine Translation
- Corpus-Based Machine Translation
 - Statistical Machine Translation
 - Example based Machine Translation(EBMT)
- Hybrid Machine Translation

Rule based machine translation is also called as Knowledge Based MT. It is considered a Classical Approach to Machine Translation. Here translation is done based on a few pre-coded rules or based on the linguistic details of the source language and the target language. In this method, the rules are defined beforehand, that involves morphological, syntax and semantics of the source and target language. There are three sub fields in RBMT – Direct, Interlingua and Transfer based MT. All the three follow the same principle with subtle differences.

Direct Machine Translation approach is one of the first MT approaches to be proposed and this makes it one of the oldest approaches. This method depends very heavily on a bilingual dictionary. Minimal amounts of analysis on the syntax and semantics are done. Every word of the source language is substituted for its corresponding word in the target language referring to a bilingual dictionary.

In Interlingua based approach, the text in the source language is converted into an intermediate representation, which is then converted into the target language. Usually we can generate text any language from the intermediate representation and hence there is no need to develop a new system for every language pair that we encounter. This is one of the major advantages of Interlingua based approach over direct MT.

Transfer based approach is similar to Interlingua based approach but one difference is that transfer based approach depends to a certain extent on the language pair that is involved.

Corpus based MT approach has two subfields namely statistical and example based approach. Corpus based approaches makes use of parallel corpus as the input to calculate the probabilities. Parallel corpus are documents that contain input language contents as well as target language content.

Statistical MT approach is based on statistical models that have been structured from the parallel corpora. Probability is the main mathematical concept that is used in this approach. We usually apply Bayes Theorem to find the probability distribution that a given word 'a' is the translation of the word 'b' ($P(a/b)$). Statistical approach is the new wave in the industry and

many companies are moving to this approach. However, one of the most important disadvantage of this method is that the corpus creation can be very costly.

In example-based approach, if have a predefined set of translations and taking these translations as reference, we do the translation of other sentences. Given a text to be translated, we choose the most similar translation from the predefined set of translations and use it for translation.

Hybrid Machine Translation combines both the rule based as well as corpus-based approaches to translate text from one language to another.

3. Literature Survey

Since the proposed MT system translates English to Hindi sentences, the literature survey focusses on the MT works, which translates text from English to Hindi.

ANGLABHARTI is a multilingual MT project, which translates English to Indian languages, primarily Hindi [6]. ANGLABHARTI is based on a pattern directed approach in the first stage; a pattern based parsing is performed on the English sentence which outputs a 'pseudo-target' that is applicable to a set of Indian languages. Word Sense Disambiguation (WSD) has been incorporated using semantic tags. The ANGLABHARTI system is currently being implemented from English to Hindi translation called AnglaHindi, which is web-enabled (<http://anglahindi.iitk.ac.in>).

The Anusaaraka system translates simple English sentences to Hindi language [7]. To analyze the English sentences, it uses paninian grammar for translation. The advantage of Anusaaraka is that, after the completion of the source text analysis, the user can lessen the complexity of the sentence in case of erroneous translation.

MaTra is a human assisted MT system that translates English to Indian languages (at present Hindi [8]). The system works based on a transfer approach that uses frame-like structured representation. The approach has tried to resolve the ambiguities that are induced by the rule-based and heuristic approaches. The system primarily works for news, annual reports, and technical phrases.

Mantra MT system focusses on translating gazette notifications related to government appointments and parliamentary proceedings mainly between English and Indian languages [9]. It uses Tree Adjoining Grammar (LTAG) formalism. It retains the format of the input language. It has been successfully tested with language pairs, namely, Hindi-English and Hindi-Bengali translation.

The proposed system puts forth a Machine Translation technique that uses the concept of N-Grams and statistics to translate various types of English sentences to Hindi. The proposed MT system

has been tested and compared with an existing rule based system developed in 2016 using the IIT Bombay corpus [10].

4. Proposed Model

The training is done by identifying by using the N gram model. An n -gram is a contiguous sequence of n words from a given sentence or text. They are a set of co-occurring words within a given limit. An n -gram of size 1 is referred to as a "unigram"; size 2 is a bigram (or, less commonly, a "di-gram"); size 3 is a trigram, size 4 is a four-gram and so on. For example, consider the sentence: Ajay and Ravi are best friends. If $N = 2$ (known as bigrams), then the n -grams would be:

- Ajay and
- and Ravi
- Ravi are
- are close
- close friends

The proposed system uses unigrams, bigrams, and trigrams. After identifying the N gram, morphological analysis is done for the input words. The bidirectional dictionary for English to Hindi is used to get the corresponding Hindi words for the English words. The most likely word order of the Target language, which is in this case, Hindi is chosen by using Naïve Bayes propobability. Next, the system uses the Naïve Bayes algorithm to calculate the order of the list of words. The Naïve Bayes algorithm give the sequence of words which has the highest word order. The Naïve Bayes posterior probability for the expected sequence is calculated as follows [11].

$$P(h/D) = \frac{P(D/h) * P(h)}{P(D)}$$

$P(h)$: Prior probability of hypothesis h

$P(D)$: Prior probability of training data D

$P(h/D)$: Probability of h given D

$P(D/h)$: Probability of D given h

The above formula can also be represented as follows:

$$posterior = \frac{prior * likelihood}{evidence}$$

4.1 Types of sentences handled

The proposed system handles, simple, compound and complex sentences. A simple sentence has only the basic elements that are needed to make a proper sentence. A compound sentence, on the other hand is made up of two independent clauses that are connected by conjunctions like and, but, or, not and many others. A complex sentence is a sentence containing a subordinate clause or clauses.

Example 1: gives a glimpse of the types of sentences handled:

How are you? (Simple sentence)

He had bread and butter (compound sentence)

Although she had her doubts, she went ahead with the plan (complex sentence)

Example 2 gives the sentence and their respective translations generated by the proposed model

There was a nest on that tree.

वहापेडपरगोसलाहै।

Can you make this work?

कियातुमइसेकामबना।

Can I help you?

क्यामैंआपकीसहायताकरसकताहूँ?

Happy New Year!

नएसालकीबधाईयाँ।

Happy birthday!

जन्मदिनमुबारकहो !

He has a beard.

उसकेपासदाढ़ीहै।

My bag was stolen.

मेराबस्ताचोरीहोगयाथा।

5. Performance Evaluation and Comparison

The proposed system has been tested with agriculture domain parallel corpus provided by IIT Bombay. The number of sentences in the corpus is 1020. The proposed system has also been tested with 240 manually collected English sentences from the web documents. The number of sentences that has been used in the training set 1,00,000.

A performance comparison has also been done with one of the existing rule-based systems that was developed in 2016. It is based on a technology that was developed by the Applied Artificial Intelligence Group of C-DAC, Pune [12].

The working of the rule-based system is as follows:

The system puts forth a simple MT framework which makes use of Parts of Speech (POS) tags of the words as features and simple decision tree to translate various types of English sentences to Hindi. The construction of the decision tree makes sure that the sentence is not treated as just as a collection of words but also preserves the grammatical relationship between different sentences the sentences. This is the most important use of the construction of the decision tree. The meaning of the sentence is preserved with the help of the POS tags. Based on these tags, the words get placed at different positions in the decision tree. In short, using the POS tags to construct the decision trees results in translations that tries to retain the meaning of the sentence perceived in the source language.

The reason for choosing this tool for comparison is that it mainly focuses on conversion of English text to Hindi. This comparison gives us a lot of scope for improvement and also gives us a clear understanding of

where our system stands and how good it is. Most importantly, it proves the fact that statistical approaches give us much better results compared to Rule-based techniques and machine translation systems can only be improved using better statistical methods. The performance is evaluated using the Recall measure which is defined as given in the Equation 1.

$$\text{Recall} = \frac{\text{Total number of correct translations}}{\text{Total number of translations}} \text{-----}(1).$$

The Recall measures for both the systems using the IIT Bombay parallel corpus is given and that of manually collected 240 sentences are given in Table 1.

Table 1. Performance Evaluation using manually collected sentences

Factors	Proposed Statistical System		Rule Based system	
	IIT Bombay Corpus	Manually Collected Corpus	IIT Bombay Corpus	Manually Collected Corpus
Total Number of Correct Translations	749	183	366	123
Total Number of Translations	1020	240	1020	240
Recall	73.43%	76.25%	35.88%	51.25%

Example 3 gives the translations given by both the systems. (The term, MT1 stands for the rule-base system and MT2 stands for the proposed system):
Example 3: Translations of both the systems

1. What is your name?

MT1: आपकेनामक्याहै? (Aapakenaamkyahai?)

MT2: आपकानामक्याहै? (Aapakanaamkyahai?)

2. How are you?

MT1: तुमकैसेहै। (Thumkaisehai.)

MT2: तुमकैसेहो?(Thumkaise ho?)

3. Can you make this work?

MT1: कियातुमइसेकामबना। (Kya tum isekambana)

MT2: कियातुमइसेकामकरवासकतेहो?(Kyatumissekamkarvasakthe ho?)

4. How is the weather today?

MT1: मौसमआजकैसाहै। (mauamaajkaisahai.)

MT2: आजमौसमकैसाहै?(Aajmausamkaisahai?)

5. There was a nest on that tree.

MT1: वहपेडपरगोसलाहै। (Vohped par ghoslahai)

MT2: उसपेडपरगोसलाहै। (Oosped par ghoslahai)

6. He snores very loudly.

MT1: वहअत्यधिकऊँचेस्वरमेंखरटिलेतेहैं! (Vah athyadhikuncheswar me kharratelethehai.)

MT2: वोबहुतजोरसेखरटिलेताहै। (Vo bahutjhor se kharratelethehai.)

7. This is an exquisite piece of art.

MT1: यहकलाकीबकायापिकहै। (Yeh kala kibakayapikhai.)

MT2: यहकलाकासुंदरटुकड़ाहै। (Yeh kala ka sundartukadahai.)

8. Unfortunately, noone was present.

MT1: दुर्भाग्यसेनहींकिसीभीकोईभीवर्तमानथा। (Dhurbhaghya se nahikisibhi koi bhi varthamantha.)

MT2: दुर्भाग्यसेएकमोजूदनहीथा। (Dhurbhagya se ekmoujudhnahitha.)

9. Mumbai is famous for its textiles.

MT1: मुंबईअपनावस्त्रकेलिएमशहूरहै। (Mumbai aapnavasthrakeliyemashuhoorhai.)

MT2: मुंबईवस्त्रोकेलिएमशहूरहै। (Mumbai vasthronkeliyemashuhoorhai.)

10. When does the train arrive?

MT1: रेलगाड़ीकबपहुँचेंगे? (Relgadikabpahunche?)

MT2: रेलगाड़ीकबपहुँचेंगा? (Relgadikabpahunchega?)

One reason why the rule-based system does not perform at par with the statistical system is because, it is not flexible. It always follows one pattern or a set of hard coded rules for the translation of sentences. Different sentences have different interpretations and hence need different rules for translating the sentences. However, the rule-based system does not follow this. It

tries to apply the same set of rules for different types of sentences and this is the main reason for its low performance when compared to the proposed system. It is not possible to generalize rules pertaining to different sentences for translation. Every sentence has a unique structure and hence for achieving a reasonable performance, we need to try to accommodate as many sentences as possible. The variety in sentences is in fact one of the most challenging hurdles in the process of translation.

Another reason for rule-based system's low performance is its property to retain the format or structure of the source language. Different languages have different structures or sentence patterns and hence we end up achieving very little by preserving the format of the source sentence. We need to preserve the relationship between the words but not the actual structure of the sentence. However, the rule-based system does not do so. It tries to replicate the source language structure into the target language. This is one big disadvantage because the sentence structure of Hindi and English differ.

6. Conclusion and Future Enhancements

Language is a form of art. Different languages have different knowledge bases that needs to be explored. This paper makes such an attempt by proposing an automatic translation of English sentences to Hindi language using simple rules. Statistical approaches are gaining popularity day by day and have shown very good success rates in the field of machine translation. One drawback of this system is its dependence of the training data. We need to make sure that our training data is very large and covers a wide range of vocabulary. Our translation heavily depends on the probability calculation (statistics) of the words and these words are got from our training data. Therefore, it is necessary that the training data used is self-sufficient. Translation systems can be widely used in defense divisions for translating text and documents from different languages. Instead of wasting work force on such tasks, we can use the machine translation systems for this purpose. This will save us a lot of time and money. Documents can be translated into the desired target language instantly (in a matter of milliseconds or in some case seconds) whereas other means might take lots of time.

In addition, this method can be improved by using other machine-learning algorithms like Decision Trees, K – Nearest Neighbors etc. instead of Naïve Bayes. Using other algorithms might improve the success rate of the system. In addition, we need to train our model on a continuous basis. One time training is not sufficient. Languages evolve continuously and hence for the system to keep up with the new trend, training regularly is very important.

References

- [1] Koehn Philipp. EuroParl: A Parallel Corpus for Statistical Machine Translation. Machine Translation Summit. (2005).
- [2] Dasgupta, S., Wasif, A. and Azam, S. An Optimal Way towards Machine Translation from English to Bengali, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT). 2004.
- [3] Singh G. A Punjabi to Hindi Machine translation system". In proceeding of: COLING, 22nd International Conference on Computational Linguistics. 2008.
- [4] Dwivedi, Sanjay K., and Pramod P. Sukhadeve. "Machine translation system in indian perspectives." *Journal of computer science* 6.10 (2010): 1111.
- [5] Tripathi, Sneha, and Juran Krishna Sarkhel. Approaches to machine translation. NISCAIR-CSIR.*ALIS Vol.57(4) .December 2010.*
- [6] Sinha, R. M. K., and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system." MT Summit IX, New Orleans, USA (2003): 494-497.
- [7] Bharati A, Chaitanya V, Kulkarni AP, Sangal R. Anusaaraka: Machine translation in stages. VIVEK-BOMBAY-. 1997.
- [8] Ananthakrishnan R, Kavitha M, Jayprasad JH, Shekhar RS, Bade SM. Matra: A practical approach to fully-automatic indicative english-hindi machine translation. InSymposium on Modeling and Shallow Parsing of Indian Languages (MSPIL'06) 2006 (pp. 1-8).
- [9] Naskar S, Bandyopadhyay S. Use of machine translation in India: Current status. AAMT Journal. 2005 Sep 13:25-31.
- [10] Jha GN. The TDIL Program and the Indian Language Corpora Initiative (ILCI). InLREC 2010 May 17.
- [11] Peng F, Schuurmans D. Combining naive Bayes and n-gram language models for text classification. InEuropean Conference on Information Retrieval 2003 Apr 14 (pp. 335-350). Springer, Berlin, Heidelberg.

[12] Darbari H. Computer-assisted translation system—an Indian perspective. In Machine Translation Summit VII 1999 (pp. 80-85).

[13] S.V.Manikanthan and V.Rama “Optimal Performance of Key Predistribution Protocol In Wireless Sensor Networks” International Innovative Research Journal of Engineering and Technology ,ISSN NO: 2456-1983,Vol-2,Issue –Special –March 2017.

[14] T. Padmapriya and V.Saminadan, “Improving Performance of Downlink LTE-Advanced Networks Using Advanced Networks Using Advanced feedback Mechanisms and SINR Model”, International Conference on Emerging Technology (ICET), vol.7, no.1, pp: 93, March 2014.

