ABSTRACT
Despite the linguistic diversity, communication is important for economic and social growth. But it is impossible for a human being to know all languages. This led to the inception of machine translation. Machine Translation (MT) is a field of Artificial Intelligence and Natural Language Processing which deals with translation from one language to another using machine translation system. Even after translation, in order to assess the goodness of MT system i.e. if the translated output is of human translation quality or not, some Evaluation strategy is required. This paper gives a review of the work done on various Indian machine translation systems and existing methods for evaluating the translated MT system’s Output.

General Terms
Machine Translation, Natural Language Processing.

Keywords
Computational Linguistics, Morphology

1. INTRODUCTION
India is home to not just one or two languages but to a myriad number of diverse linguistic families. Mainly the languages belong to two chief families - Indo Aryan and Dravidian. These are spoken by more than 90% of Indian population. Apart from these two, Austro-Asiatic languages and Tibeto-Burman linguistic languages also give key contribution to the language list.

More than 30 languages and approximately 2000 dialects are used for communication in India, amongst which Hindi and English are taken as languages for official work. There are 22 scheduled languages for different states which include Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Kannada, Kashmiri, Konkani, Maithili, Santali, Sindhi, Tamil, Telugu, and Urdu. [1]

It is difficult for a human being to know all these languages and hence the need for a translator arises, where the translation from one language to another is done. Manual translation of these language pairs is a very cumbersome task and thus we opt for the automatic machine translation systems, wherein computer software translates one natural language to another which the human understands.

The history of machine translation dates back to July 1949 when Warren Weaver (a director at the Rockefeller Foundation, New York) wrote an influential paper which introduced Americans to the idea of using computers for translation. The first conference on MT came in 1952. There was the first demonstration of a translation system in January 1954, and it attracted a great deal of attention and since then there has been no stopping. The field of Machine Translation has been expanding limitlessly in all countries including India.

Machine translation is useful in many places where: 1) Highly repetitive content is present 2) Content is similar to translation memories but not exactly the same 3) High value content that is changing every hour and every day and there is time sensitivity 4) Content does not need to be perfect but just approximately understandable. And many other areas.

2. APPROACHES OF MACHINE TRANSLATORS
Many technical approaches have been developed to solve the challenges of language translation. Some of these approaches include:

2.1 Rule Based Machine Translation (RBMT)
Commonly known as “Knowledge-Based Machine Translation, is based on linguistic information about source and target languages. RBMT systems are basically constituted by two components: the rules that account for the syntactic knowledge, and the lexicon, which deals with the morphological, syntactic, and semantic information of the language. [2] These retrieved from dictionaries and grammars of each language. RBMT system translates input sentence (in source language) to output sentences (in some target language) on the basis of rules and lexicons of both the source and the target languages involved in translation. There are three different types of rule-based machine translation systems:

2.1.1. Direct Systems
The Dictionary Based Machine Translation maps output and input using basic rules.

2.1.2. Transfer based RBMT Systems
The Transfer Based Machine Translation uses syntactic and morphological features of language (lexicons).

2.1.3. Interlingual RBMT Systems
Here source language is transformed into an intermediate language which is independent of any of the languages involved in the translation. This intermediate representation is known as Interlingua, which can be transformed into multiple languages. [3][4]
2.2 Example Based Machine Translation (EBMT)

This approach uses a bilingual corpus for translating one language to another. EBMT matches the sequence of words with words in corpus by decomposing the complete sentence into fragments and matching these fragments against proper examples in the corpus. It uses analogy translation principle.

The performance of this method depends on the corpus. The key things about corpus include size of the corpus (number of samples in corpus), quality of corpus and also if the corpus is supervised, unsupervised or semi-supervised.

2.3 Statistical Machine Translation (SMT)

Warren Weaver in 1949 introduced the idea of Statistical Machine Translation. In SMT, translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. This means that we apply a learning algorithm to a large body of previously translated text, known variously as a parallel corpus, parallel text, bitext, or multilingual.

2.4 Hybrid Machine Translation System

Statistical and rule-based MT complements each other. One overcomes shortcomings of the other, with their very different strengths and weaknesses. An optimized MT architecture should include elements of both theories. Hence the Hybrid Machine Translation System comes into picture.

Hybridization of machine translation architectures can be done using various methods: 1) Hybridization guided by RBMT 2) Hybridization guided by corpus-based MT.

3. INDIAN MACHINE TRANSLATION SYSTEMS

The table below introduces existing MT systems in India based on different translation Approaches (explained in earlier section) along with the year of release, key people involved, language pair and details regarding the system:

<table>
<thead>
<tr>
<th>#</th>
<th>Translation System</th>
<th>Year</th>
<th>People Responsible</th>
<th>Source Language</th>
<th>Target Language</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>A)</td>
<td>Direct Machine Translation Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Anusaaraka systems among Indian Languages [28][29]</td>
<td>1995</td>
<td>Rajeev Sangal</td>
<td>Telugu, Kannada, Bengali, Punjabi and Marathi</td>
<td>Hindi</td>
<td>The output of the system followed the grammar of the source language only. Developed by IIT Kanpur (earlier), IIT Hyderabad (Now)</td>
</tr>
<tr>
<td>2.</td>
<td>Punjabi to Hindi MT System [30][31]</td>
<td>2007-2008</td>
<td>G S Josan and G S Lehal</td>
<td>Punjabi</td>
<td>Hindi</td>
<td>Based on direct word-to-word MT approach. Accuracy of this system is 90.67%. Developed by Punjabi University, Patiala.</td>
</tr>
<tr>
<td>4.</td>
<td>Hindi-to-Punjabi MT System [12][17][18][19]</td>
<td>2009-2011</td>
<td>Goyal V and Lehal G S</td>
<td>Hindi</td>
<td>Punjabi</td>
<td>The translation accuracy of the system is 87.60% on the basis of accuracy test. Developed by Punjabi University, Patiala.</td>
</tr>
<tr>
<td>B)</td>
<td>Transfer-Based MT Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>MANTRA MT [13][21]</td>
<td>1999</td>
<td>Hemant Darbari and Mahendra Kumar Pandey</td>
<td>English</td>
<td>Hindi, Bengali, Telugu, Gujarati</td>
<td>Translates in specific domain of personal administration that includes gazette notifications, office orders, office memorandums and circulars Uses TAG and LTAG to represent English &amp; Hindi grammar. It is based on synchronous Tree Adjoining Grammar and uses tree transfer for translating from English to Hindi.</td>
</tr>
<tr>
<td>No.</td>
<td>System</td>
<td>Year</td>
<td>Languages Used</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------------------------------------</td>
<td>-------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>An English–Hindi Translation System</td>
<td>2002</td>
<td>Gore L and Patil N</td>
<td>Uses different grammatical rules of source and target languages and a bilingual dictionary for translation. The domain of the system was weather narration.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>MAT</td>
<td>2002</td>
<td>Murthy K</td>
<td>Uses UCGS(Universal Clause Structure Grammar), morphological analyser &amp; post-editing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C)</td>
<td>Interlingua Machine Translation Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Machine Translation Systems</td>
<td>Year</td>
<td>Authors</td>
<td>Source of Data</td>
<td>Language Combination</td>
<td>Translation Approach</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------</td>
<td>------</td>
<td>---------</td>
<td>----------------</td>
<td>---------------------</td>
<td>---------------------</td>
</tr>
</tbody>
</table>

**D) Hybrid Machine Translation Systems**

<table>
<thead>
<tr>
<th></th>
<th>System</th>
<th>Year</th>
<th>Authors</th>
<th>Source of Data</th>
<th>Language Combination</th>
<th>Translation Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Anubharti Technology [168]</td>
<td>1995, 2004</td>
<td>Sinha</td>
<td>Hindi</td>
<td>Indian Languages</td>
<td>A combination of example-based, corpus-based approaches and some elementary grammatical analysis</td>
</tr>
<tr>
<td>2.</td>
<td>ANUBHARTI-II [168]</td>
<td>2004</td>
<td>R M K Sinha</td>
<td>Hindi</td>
<td>Indian Languages</td>
<td>Uses Generalized Example-Base (GEB) along with Raw Example-Base (REB) MT approach for hybridization</td>
</tr>
<tr>
<td>4.</td>
<td>Lattice Based Lexical Transfer in Bengali Hindi MT Framework [171]</td>
<td>2011</td>
<td>Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, and Anupam Basu</td>
<td>Bengali</td>
<td>Hindi</td>
<td>Uses transfer based MT approach with the help of lattice-based data structure</td>
</tr>
</tbody>
</table>

**E) Example Based Machine Translation (EBMT) Systems**

<table>
<thead>
<tr>
<th></th>
<th>System</th>
<th>Year</th>
<th>Authors</th>
<th>Source of Data</th>
<th>Language Combination</th>
<th>Translation Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>ANUBAAD [172]</td>
<td>2000, 2004</td>
<td>Bandyopadhyay S</td>
<td>English</td>
<td>Bengali</td>
<td>Domain specific to English Headlines translation Example-base, Generalized Tagged example-base and Phrasal example-base are separately maintained If the headline cannot be translated using above methods then the heuristic translation strategy is used</td>
</tr>
<tr>
<td>5.</td>
<td>Hinglish machine translation system [178]</td>
<td>2004</td>
<td>Sinha and Thakur</td>
<td>Hindi</td>
<td>English</td>
<td>Based on AnubBarti-II and AnglaBharti-II Performs very shallow grammatical analysis</td>
</tr>
<tr>
<td>6.</td>
<td>English to (Hindi, Kannada, Tamil) and Kannada to Tamil Language-Pair Example Based MT [179,180]</td>
<td>2006</td>
<td>Balajapally P., P Pydimarri, M Ganapathiraju, N Balakrishnan and R Reedy</td>
<td>English and Kannada</td>
<td>Hindi, Kannada and Tamil</td>
<td>Based on a bilingual dictionary comprising of sentence dictionary, phrases dictionary, words dictionary and phonetic dictionary.</td>
</tr>
</tbody>
</table>
7. The MATREX System [41][42] 2008 Ankit Kumar Srivastava, Rejwunul Haque, Sudip Kumar Naskar and Andy Way English Hindi Uses marker based chunking and “edit-distance style” dynamic programming alignment algorithm Domain limited to Conference papers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2. English to Indian Languages Machine Translation System[27]</td>
<td>2006</td>
<td>Consortium of Nine institutions</td>
<td>English</td>
<td>Indian Languages</td>
<td>Limited to Tourism and Healthcare domain Uses statistical techniques and tools including the POS tagger, parser, decoder</td>
</tr>
</tbody>
</table>

* C-DAC Mumbai, IISc Hyderabad, C-DAC Pune, IIT Mumbai, Jadavpur University Kolkata, IIIT Allahbad, Utkal University Bangalore, Amrita University Coimbatore and Banasthali Vidyapith Banasthali

4. EVALUATION OF MACHINE TRANSLATION

There are many systems in existence for translation but translation merely is not sufficient, it should be understandable, acceptable and must be of good quality. Hence, in order to judge the quality of translation, some evaluation measures are required.

The main aim of machine translation evaluation is to check that how well the machine’s translated output correlates with human’s reference translated output, for same language MT Evaluation strategies were initially proposed by Miller and Beeber-center in 1956 followed by Pfaffline in 1965. In the beginning MT evaluation was carried out only by human judges. This process, however, was time-consuming and highly subjective. Then as the field of machine translation grew there arose the dire need for automation i.e., for fast, objective, and reusable methods of evaluation, the results of which are not biased or subjective at all. To this date, several metrics for automatic evaluation have been proposed and which are accepted by the MT community enthusiastically, but the research is never ending.


4.1 Human Evaluation

Manual evaluation is done by calculating fluency, adequacy and fidelity (Hovy, 1999; White and O’Connell, 1994). Adequacy is used to evaluate the quantity of the information existing in the original text that a translation contains. Commonly fluency refers to the degree to which the translation is well-formed according to the grammar of target language [8]. Fidelity refers to the amount of information retained in translated output in comparison to candidate.

In human evaluation there are two types of evaluators: Bilingual, those who understand both source and target language and Monolingual i.e. understanding only target language.

Here, the human evaluator looks at the translation and judges it to check that if it is correct or not based upon factors described above. The score of human evaluator is given on a particular scale based on which the translations are ranked.

One of the most important peculiarities of human evaluation is that two human evaluators when judging the same text could give two different evaluations, as might the same evaluator at different moments (even for exact matches). Which means that human criteria for evaluation of Machine output is subjective. Also human evaluations are non-reusable, expensive and time consuming. To overcome these situations we need automatic system which can perform faster and give the output if not same but at least comparable to human output and can be reused over and over.

4.2 Automatic Evaluation

Human Evaluations are actually gold standards but the main issue in such evaluation is Cost and Time. Humans take more time and are expensive. Hence we need automatic metrics which are: 1. Quick 2. Inexpensive 3. Language-independent 4. Correlate highly with human evaluation 5. Have little marginal cost per run [9].

Mostly all automatic metrics are based on either Edit Distance Based, Precision Based, Recall Based, F-measure based. The boom of automatic metric started with the introduction of BLEU (Papineni et al., 2001) which is based on average of matching n-grams between candidate and reference. Following IBM’s lead NIST (Doddington, 2002) came out, which calculates matched n-grams of sentences and attach different weights to them. GTM (Turian et al., 2003) computes precision, recall and f-measure in terms of maximum unigram matches. In same year ROUGE (Lin and Hovy, 2003) was introduced that created the summary & compared it with the summary created by human (Recall oriented). 2005 proved to be very important because one of the most successful metric METEOR (Banerjee & Lavie, 2005)
Use of...

2) "MAT: A Machine Assisted Translation System", Vineet Chaitanya, Rajeev Sangal, Ankush Gupta, Sitender, Seema Bawa, "Survey of Indian Machine Translation Systems in India along with the approaches used for translation (i.e. rule-based, hybrid and statistical approaches). Not many resources are available for free word order languages, morphologically rich languages and Resource poor languages and most of the translators are domain specific which focus only on particular domain translations. Also we discussed evaluation strategies for evaluating the translated output of machines. Many Human evaluation strategies have been applied and various automatic methods of evaluation (Metrics) have also been proposed off-late, to assess the quality of translation, but there is still no metric in existence which can perform remarkably well for all the languages at one time i.e. it is not comparable to human assessment. Hence MT is an open research field even today.

5. CONCLUSIONS

In this paper we did a brief survey of existing Machine Translation Systems in India along with the approaches used for translation (i.e. rule-based, hybrid and statistical approaches). Not many resources are available for free word order languages, morphologically rich languages and Resource poor languages and most of the translators are domain specific which focus only on particular domain translations.

6. REFERENCES


