Improving Translation Quality By Using Ensemble Approach

Deepti Chopra
Department of Computer Science
Banasthali University
Newai, India
deeptichopra11@yahoo.co.in

Nisheeth Joshi
Department of Computer Science
Banasthali University
Newai, India
jnisheeth@banasthali.in

Iti Mathur
Department of Computer Science
Banasthali University
Newai, India
mathur.iti@rediffmail.com

Abstract—**Machine translation (MT) has been a topic of great research during the last sixty years, but, improving its quality is still considered an open problem. In the current paper, we will discuss improvements in MT quality by the use of the ensemble approach. We performed MT from English to Hindi using 6 MT different engines described in this paper. We found that the quality of MT is improved by using a combination of various approaches as compared to the simple baseline approach for performing MT from source to target text.**

Keywords-machine translation; named entity translation; natural language processing; source text rewriting

I. INTRODUCTION

Machine translation is a natural language processing (NLP) application. It is defined as the process of conversion of text from one language to another and is still considered an open problem. Tasks related to MT began soon after the World War II, when translation was performed with the help of electronic bilingual dictionaries and manually designed lexical rules [1]. To make advancements in the field of MT, US Government established a committee called Automatic Language Processing Advisory Committee (ALPAC). ALPAC members concluded that MT was not very accurate, and was more expensive than human translation. So, they suggested investing in basic research in NLP. Recently, more powerful computers were developed that could handle the huge amount of the MT related data. Today, when we look back into past, we may realize that the ALPAC report led to progress in the field of NLP in the long term. Several NLP based resources have been developed and helped developers to solve MT based problems. Today, a large number of companies and institutions have been motivated by the profitability of MT as a business and this has led them to invest in MT based projects.

II. CLASSIFICATION OF MT SYSTEMS

On the basis of degree of human interaction, MT systems can be classified into three types [2]: (a) machine-aided human translation (MAHT), (b) human-aided machine translation (HAMT) and (c) fully automatic machine translation (FAMT). MAHT is implemented by many commercial systems. FAMT systems are mostly free and can be found on the internet. According to the levels of linguistic analysis, MT may be classified into three types: (a) direct, (b) transfer and (c) interlingua. The levels of linguistic analysis can be seen in [3].

In direct approach, phrase by phrase or word by word translation takes place without undergoing any other additional representation [4]. The advantage of direct MT approach is that translation can be understood with little effort. Its disadvantages are that it can be built only for specific language pairs and it is expensive to build in case of multilingual scenarios. Also, some of the meanings of the source text might get lost in translation when using it. Example based MT (EBMT) systems and statistical MT (SMT) systems are based on direct approach. In EBMT systems, bilingual corpus or parallel texts are used. They are implemented using case based reasoning methodology of machine learning. Statistical machine translation (SMT) systems make use of Bayes decision rule and statistical decision theory in order to reduce the number of decision errors. Rule based MT (RBMT) systems are based on transfer based approach. RBMT systems can either be syntactic transfer based MT systems or semantic transfer based MT systems, in which the source text is firstly converted to source abstract representation which is then converted to target abstract representation using linguistic rules. The target abstract representation is then finally transformed into target text. In syntactic transfer based systems, source parse tree is constructed from the source text which is then transformed into target parse tree. The target parse tree is then converted into target sentence. In semantic transfer based MT systems, the source text is converted into source semantic abstract representation which is then converted into target semantic abstract representation. Target semantic abstract representation is then transformed to syntactic structure which is finally converted to target text. An advantage of transfer based approach is that it can handle ambiguities that are transferred from one language to another. One disadvantage of the transfer based approach is that the original meaning of the source text may get lost during translation. In interlingua based MT systems, a language independent based abstract representation is constructed from the source sentence which is then converted to target sentence. Source sentences in different languages having the same meaning have the same abstract representation in interlingua based MT systems. This minimizes transfer generation burden.

III. Problems in MT Regarding Indian Languages

Indian languages are free word order and morphologically rich languages. Some of the problems faced in MT in Indian languages are presented below.

A. Complex Sentences Are Not Translated Correctly

Source sentences that are complex are usually translated incorrectly. For example, consider this source text: “The Taj Mahal is one of the wonders of the world located on the south bank of the Yamuna river in the Indian city of Agra”. The above mentioned complex source text can be simplified or rewritten to obtain the following simplified source text: “The Taj Mahal is one of the wonders of the world. It is located on the south bank of the Yamuna river in the Indian city of Agra”.

B. Named Entities Are Not Identified Correctly.

Named Entity recognition (NER) should be performed prior to MT. So, that named entities are correctly identified and spelled (translated or transliterated) correctly.

IV. Methodology

MT Systems are constructed using different combinations of ensemble techniques that include classifier based approach, source text rewriting and named entity translation. MT systems that we have designed are summarized in Table I.

TABLE I. MT SYSTEMS FOR ENGLISH TO HINDI TRANSLATION

Engine No.	Scheme
M1	English-Hindi baseline system
M2	Classifier based approach incorporated in English-Hindi SMT
M3	Source text rewriting approach incorporated in English-Hindi SMT
M4	English name entity translation system incorporated in English-Hindi SMT
M5	Classifier based approach coupled with English name entity translation system and incorporated in English-Hindi SMT
M6	Source text rewriting approach coupled with English name entity translation system and incorporated in English-Hindi SMT

We have used a testing file consisting of 1100 sentences in English. In M4, named entity recognition using Stanford NER is used to detect named entities from the English text. These named entities are translated into Hindi and sentences comprising of named entities in Hindi and non-named entities in English are produced. These English-Hindi mixed sentences are sent to statistical MT for complete translation into Hindi. In M6, at first, sentence reordering is performed using classifier based approach. These reordered sentences are sent to Stanford NER for named entity recognition. These named entities are translated into Hindi and then complete translation into Hindi is performed using statistical MT. For human evaluation, we used HEval evaluation metric [5]. The language linguistic features that have been included in human evaluation metrics are:

- Translation of gender and number of nouns.
- Translation of tense in the sentence.
- Translation of voice in the sentence.
- Identification of proper noun(s).
- Use of adjectives and adverbs corresponding to nouns and verbs.
- Selection of proper words/synonyms (lexical choice).
- Sequence of phrases and clauses in the translation.
- Use of punctuation marks in the translation.
- Fluency of translated text and translator’s proficiency.
- Maintaining the semantics of the source sentence in the translation.
- Evaluating the translation of source sentence (with respect to syntax and intended meaning).

In order to assess the quality of translation, a five point scale is employed as shown in Table II.

TABLE II. DESCRIPTION OF 5 POINT SCALE IN HUMAN EVALUATION

Score	Meaning
4	Ideal
3	Perfect
2	Acceptable
1	Partially Acceptable
0	Not Acceptable

The overall score is computed for all the linguistic features using (1):

$$\text{Overall Score} = \frac{\sum \text{Score of ith feature}}{4 * (\text{Total no.of applicable features})} \qquad (1)$$

This score is also compared with adequacy and fluency score. Adequacy and fluency are represented in Tables III and IV respectively.

TABLE III. DESCRIPTION OF ADEQUACY ON 5 POINT SCALE

Score	Meaning
5	Complete Information
4	Most Information
3	Much Information
2	Little Information
1	None

TABLE IV. DESCRIPTION OF FLUENCY ON 5 POINT SCALE

Score	Meaning
5	Ideal
4	Good
3	Non Native
2	Disfluent
1	Incomprehensible

V. Results

We have used 1100 sentences for the mentioned 6 MT engines and these sentences were distributed among 10 documents having 110 sentences each. The combined document total score for all 6 MT Engines is shown in Table V. The value in bold represents the highest overall score attained by the MT Engine. Out of 10 documents, M6 has attained the highest overall score in 8 documents. The overall accuracy of MT systems is shown in Figure 1. M6 has attained the highest overall accuracy of 0.913.

TABLE V. DOCUMENT WISE OVERALL SCORE OF MT ENGINES

	M1	**M2**	**M3**	**M4**	**M5**	**M6**
DOC 1	0.42456	0.53451	0.79728	0.67322	0.8027	**0.91302**
DOC 2	0.47223	0.64324	0.8685	0.6827	0.79626	**0.87742**
DOC 3	0.45623	0.57187	0.8732	0.64068	0.7837	**0.89334**
DOC 4	0.54163	0.65432	0.88698	0.7396	0.80064	**0.8899**
DOC 5	0.43274	0.57435	0.89732	0.73058	0.8045	**0.91114**
DOC 6	0.56847	0.62341	0.96182	0.74444	0.84218	**0.96932**
DOC 7	0.57324	0.62156	**0.95778**	0.71096	0.8334	0.91216
DOC 8	0.54628	0.68942	**0.9392**	0.73568	0.84058	0.9183
DOC 9	0.53404	0.75842	0.92698	0.7437	0.84706	**0.9294**
DOC 10	0.54231	0.74351	0.93215	0.72451	0.82568	**0.91245**

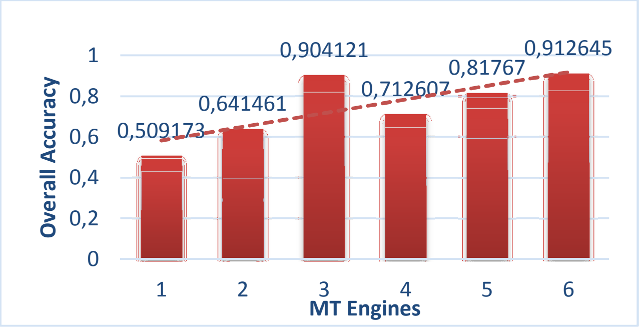

Fig. 1. Overall MT systems accuracy.

VI. CONCLUSION

In this research paper, we showed that using ensemble techniques, the quality of English to Hindi MT improves. We have designed 6 MT systems and performed our experiment on 1100 English sentences. The MT engine designed using source text rewriting approach coupled with English name entity translation system and incorporated in English-Hindi SMT has shown the highest overall accuracy of 0.913.

REFERENCES

[1] R. Srivastava, R. A. Bhat, "Transliteration systems across indian languages using parallel corpora", 27th Pacific Asia Conference on Language, Information, and Computation, pp. 390-398, Taiwan, November 21-24, 2013

[2] V. H. Yngve, "The machine and the man", Mechanical Translation, Vol. 1, No. 2, pp. 20-22, 1954

[3] B. Vauquois, "A survey of formal grammars and algorithms for recognition and transformation in machine translation", IFIP Congress (2), Vol. 68, pp. 1114-1122, UK, August 5-10, 1968

[4] W. Weaver, "Translation", in: Machine Translation of Languages, pp. 15-23, 1955

[5] N. Joshi, I. Mathur, H. Darbari, A. Kumar, "HEval: Yet another human evaluation metric", International Journal on Natural Language Computing, Vol. 2, No. 5, pp. 21-36, 2013