AN ADVANCED APPROACH FOR RULE BASED ENGLISH TO BENGALI MACHINE TRANSLATION

Chandranath Adak

Dept. of CSE, University of Kalyani, West Bengal-741235, India
adak32@gmail.com

ABSTRACT

This paper introduces an advanced, efficient approach for rule based English to Bengali (E2B) machine translation (MT), where Penn-Treebank parts of speech (PoS) tags, HMM (Hidden Markov Model) Tagger is used. Fuzzy-If-Then-Rule approach is used to select the lemma from rule-based-knowledge. The proposed E2B-MT has been tested through F-Score measurement, and the accuracy is more than eighty percent.

KEYWORDS

F-Score Measurement, Machine Translation, Rule Based Machine Translator

1. INTRODUCTION

'Machine Translation'[1-4]is the conversion procedure from one natural language to another. Now a days, it is a very challenging problem in the field of computational linguistics[5-6] and Natural Language Processing(NLP)[7]. There are thousands of languages throughout the world, and it is quite tough to know all the languages, but using machine translation, one can easily convert the unknown language information to the known one to him. For these reason, this research field has the sky-scraping demand.

We are taking two natural languages: English and Bengali [8](bi-linguistic model[9]) for rule based machine translation. Thoughthere are more than 230 millions speakers of Bengali language all over the world (mainly, Bangladesh and north-eastern part of India including West Bengal, Tripura, Assam, Bihar, and Orissa), there are limited number of resources for these language.
Here, an advanced approach of machine translation from English to Bengali language is proposed which is based on some prior knowledge based rules; and these rules are applied using fuzzy-if-then-rule approach.

2. PROPOSED METHODOLOGY

The approach for rule based E2B machine translation is as follows:

2.1. Corpus

A raw corpus is needed for any multilingual Machine Translation (MT). Here we have collected the most common Bengali words and their English term and made the *aligned parallel multilingual corpus*.

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 01–09, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3201

2.2. Tag sets for English

We have used *Penn Treebank* part of speech (POS) tags[7,10-11], eg. {Noun, Pronoun, Adjective, Verb, Adverb, Preposition, Conjunction, Interjection, Auxiliary, Determiner}≡ {NN, PP, JJ, VB, RB, IN, CC, UH, AUX, DT }. In *slash notation*, a sentence is like this:**I/**PP **need/**VB **a/**DT **pen/**NN.

2.3. Simple N-grams

A complete word string is $\{w_1, w_2, w_3, \ldots, w_n\}$ or w_1^n. The probability of occurrence of each word (considered as independent event) in its correct location:

$$P(w_1, w_2, w_3, \ldots, w_n) = P(w_1^n)$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1^2) \ldots P(w_n|w_1^{n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

The generalized N-gram approximation: $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$

For bi-gram: $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$

We can train N-gram model by counting and normalizing:

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}$$, where c = count

For bi-gram: $P^*(w_n|w_{n-1}) = \frac{C(w_{n-1} w_n)}{C(w_{n-1})}$

2.4. Smoothing

We have used the simple smoothing technique, i.e. add-one smoothing[7,12].

For N-gram: $P^*(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)+1}{C(w_{n-N+1}^{n-1})+V}$, where v =vocabulary

For bi-gram : $P^*(w_n|w_{n-1}) = \frac{C(w_{n-1} w_n)+1}{C(w_{n-1})+V}$

2.5. Stochastic PoSTagging : HMM Tagger

HMM (Hidden Markov Model) Tagger[7,13] follows a particular stochastic tagging algorithm, i.e. “pick most-likely tag” approach.

$$C_{HMM} = \max (P(w \mid t) * P(t \mid prev_{n\ tag})) ,$$

where C_{HMM} : HMM Tagger Choice, w: word, t: tag, $prev_{n\ tag}$: previous n tags.

$T=\{t_1, t_2, t_3, \ldots, t_n\}$ is the set of probable sequence of Tags in the sequence of words W.
According to the probabilistic chain rule:

$$P(T)P(W|T) = \prod_{i=1}^{n} P(w_i|w_1t_1 \dots w_{i-1}t_{i-1}t_i) \mid P(t_i|w_1t_1 \dots w_{i-1}t_{i-1})$$

Using most recent two tags approximation :

$$P(t_i|w_1t_1 \dots w_{i-1}t_{i-1}) = P(t_i|t_{i-2}t_{i-1})$$

We have to choose the tag sequence that maximizes:

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}t_{i-1})\left[\prod_{i=1}^{n} P(w_i|t_i)\right]$$

2.6. Rule-Based Model

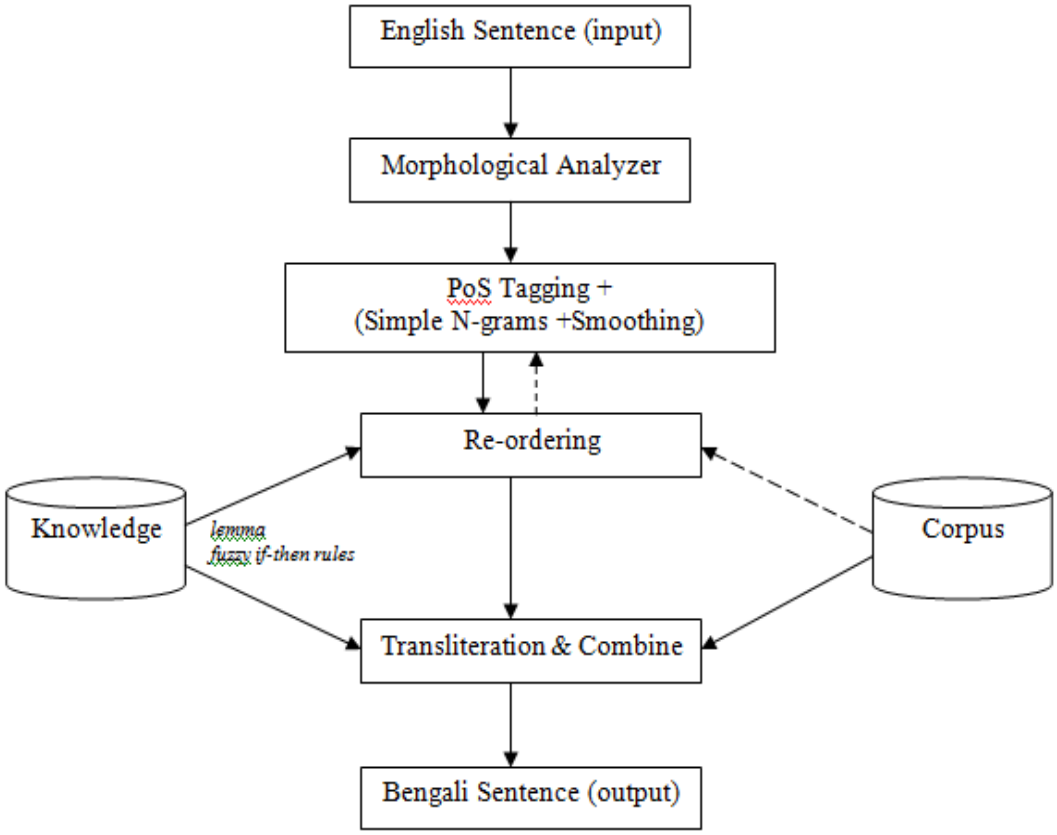

Fig.1: Proposed Rule Based Model

The English sentence is fed as input; the morphological analyzer analyzes that sentence and extracts the morphemes (smaller meaning bearing units). The morphemes are categorized in two classes :*stems* and *affixes*. Each stem is tagged with respect to tag-set of tag library (POS tagging). Now, for E2B translation, the morphemes are checked through simple N-grams and smoothing technique; the morphemes are re-ordered as and when required using knowledge based lemma and fuzzy if-then rules. Each English morpheme is converted into Bengali morpheme using our previously made corpus,these Bengali morphemes are combined together using the prior knowledge based rule to produce correct Bengali sentences as output.

2.6.1. Knowledge Based Rules

Some examples of knowledge based rules, which are used in prescribed method are as follows:

- the boy → the/DT + boy/NN → boy/NN + the/DT →ছেলে + টি→ছেলেটি
  i.e. DT + NN → NN + DT

- after morning →after/IN + morning/NN → morning/NN + After/IN →সকাল + পরে→সকাল + -এর + পরে→সকালেরপরে
  eg. before evening, under table *etc.* , i.e. IN + NN→ NN + IN

- I have a pen → I/PP + have/VB+ a/DT + pen/NN →আমি + আছে + একটি + পেন→আমার + আছে + একটি + পেন→আমার + একটি + পেন + আছে

- give me a book → give/VB + me/PP + a/DT + book/NN → me/PP + a/DT + book/NN + give/VB →আমাকে + একটি + বই + দেওয়া→আমাকে + একটি + বই + দাও

- I want a toy→ I/PP + want/VB + a/DT + toy/NN→আমি + চাওয়া + একটি + খেলনা→আমি + চাই + একটি + খেলনা→আমি + একটি + খেলনা + চাই

- I eat rice → I/PP + eat/VB + rice/NN → I/PP + rice/NN + eat/VB →আমি + ভাত + খাওয়া→আমি + ভাত + খাই
  eg. I play cricket, She drinks water *etc.* , i.e. PP + VB + NN → PP + NN + VB

- What is your name? → What/WP + is/AUX + your/PP + name/NN + ?/SP → your/PP + name/NN + What/WP + is/AUX + ?/SP →তুমি + নাম + কী + হয় + ? →তোমার + নাম + কী + হয় + ? →তোমার + নাম + কী + ?
  [SP is Sentence-final Punctuation].

- Where are you going? → Where/WRB + are/AUX + you/PP + going/VB-ing + ? /SP→ you/PP + Where/WRB + are/AUX + going/VB-ing + ? /SP →তুমি + কোথায় + যাওয়া-ইছ + ? →তুমি + কোথায় + যাচ্ছো + ?

3. EXPERIMENTAL RESULTS

The followings(*fig.2(a)-(b)*) are the screenshots of the SQL database table:

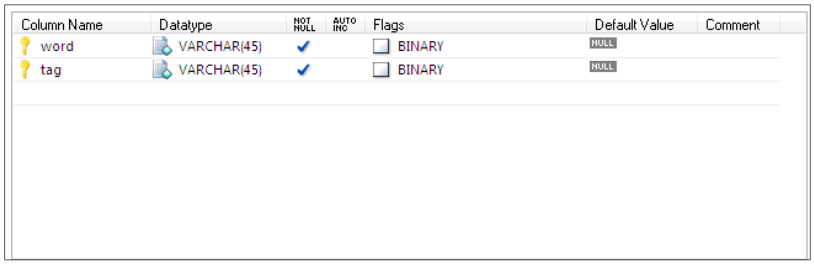

Fig.2(a): Table 'taglib', which stores the English word and its equivalent PoS Tag

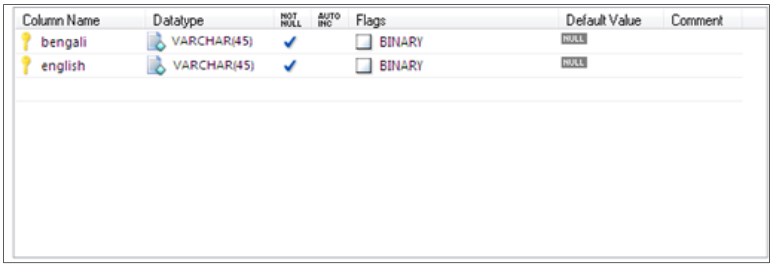

Fig.2(b): Table 'corpus' stores the English and its equivalent Bengali words

We are usingJAVAX-Swing for implementing our E2B Machine Translator(MT).The packages and classes are shown in *fig.3*.

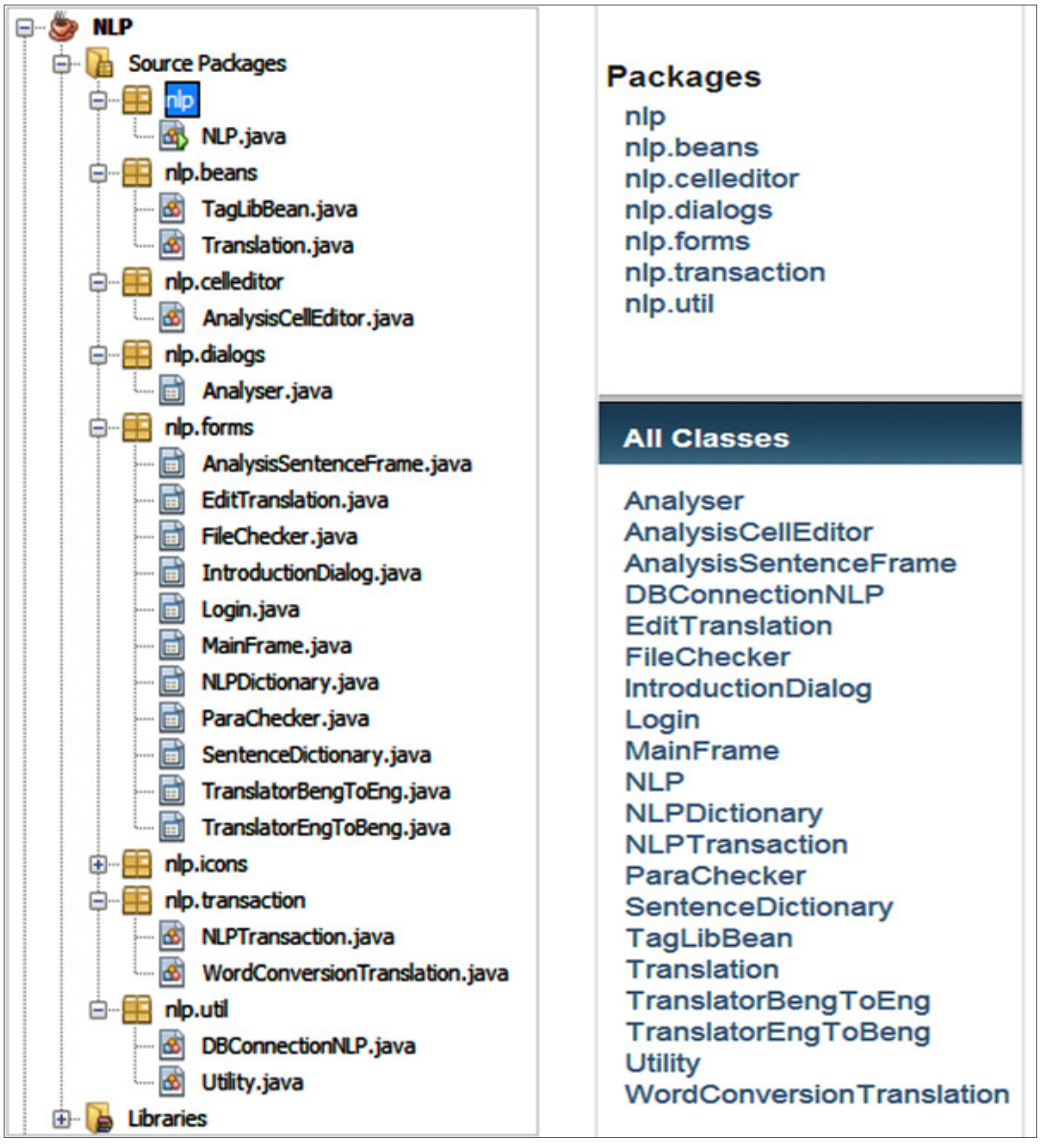

Fig.3:Packages and classes of Proposed E2B MT using JAVAX Swing

Some GUI screenshots of E2B MT are as follows (*fig.4(a) – (e)*):

Fig.4(a) : English to Bengali Translation in E2B MT

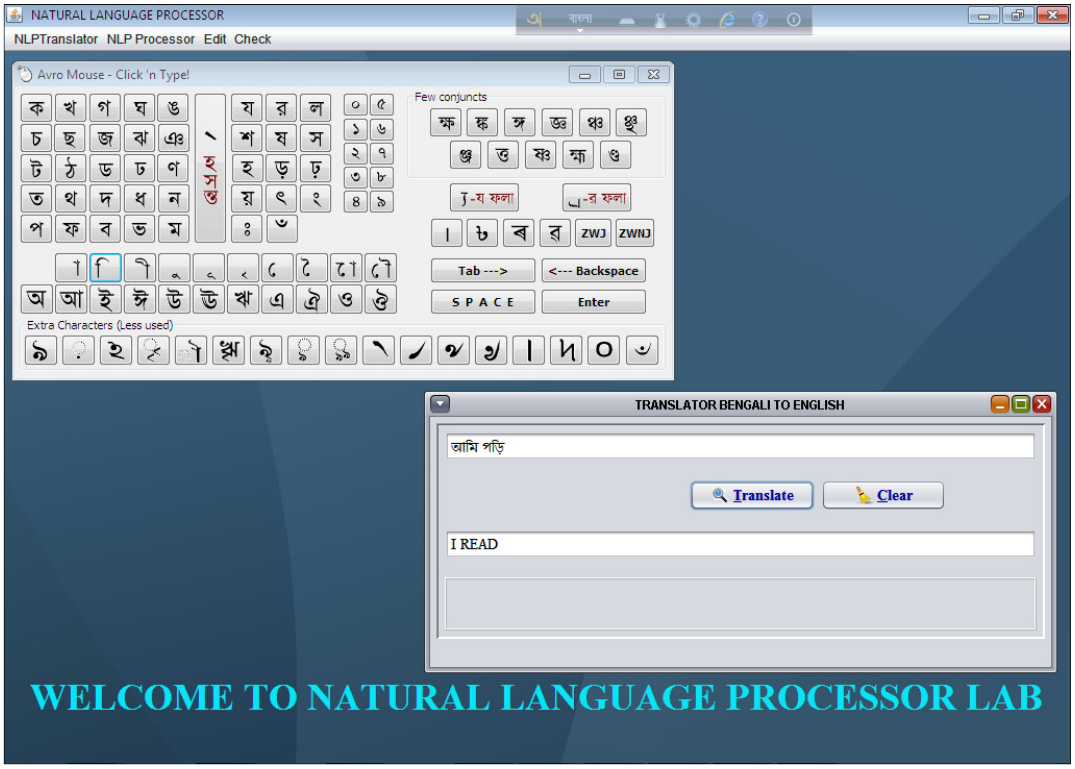

Fig.4(b) : Bengali to English Translation (An Extended part of E2B MT),

AVRO-Keyboard is used for Bengali letter typing

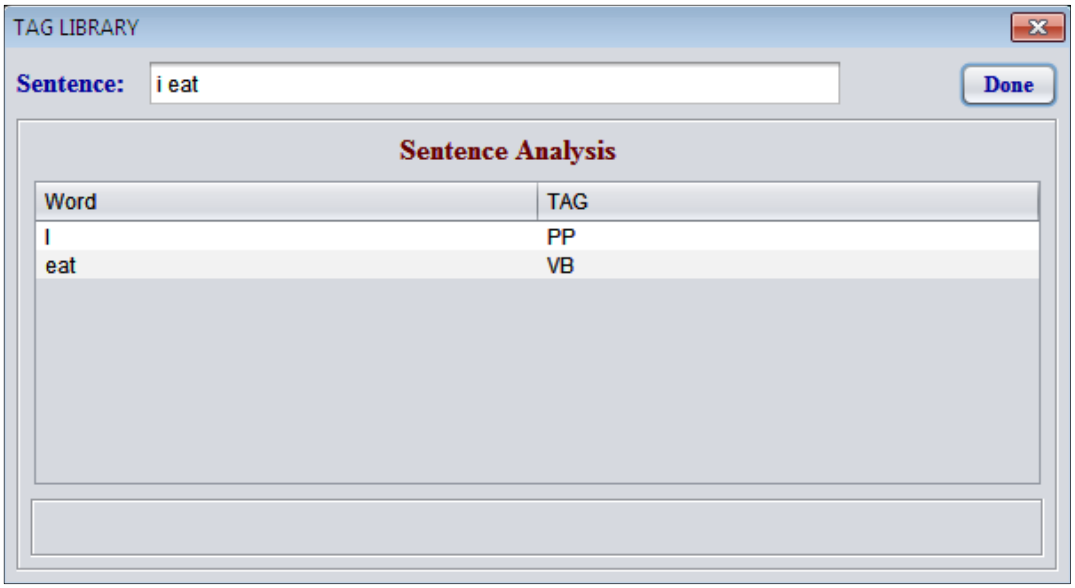

Fig.4(c) :PoS Tag Set for English Words

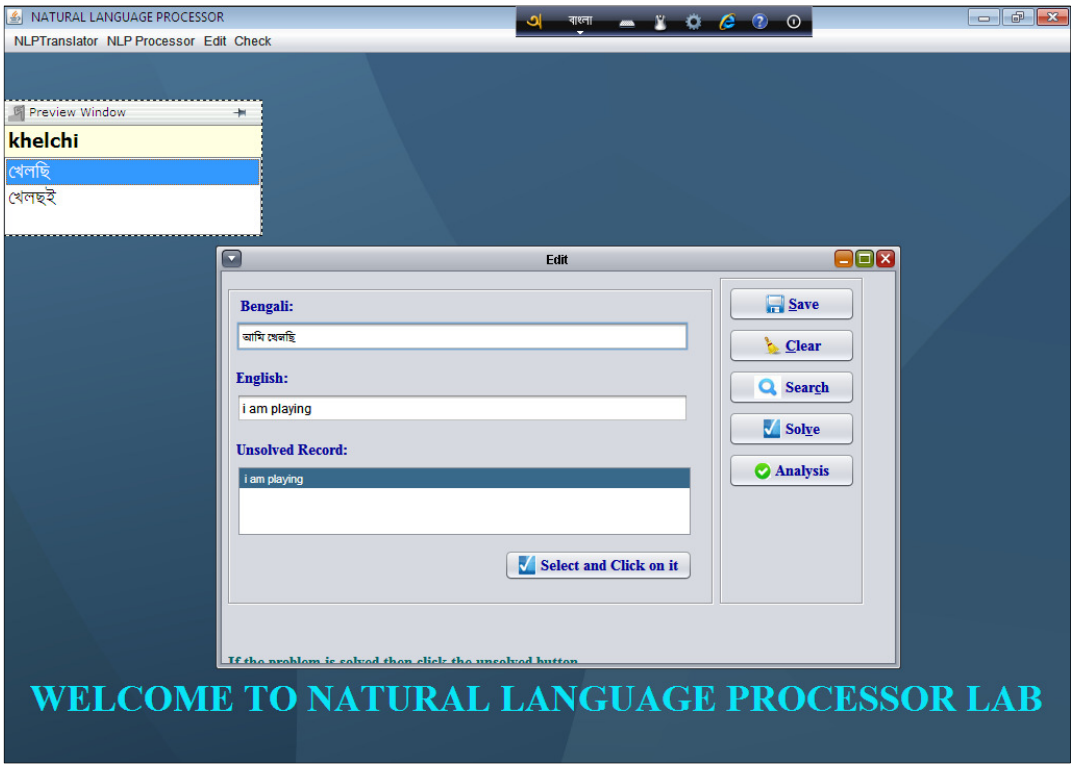

Fig.4(d) : Some extended features (*eg.* save, search, solve, analysis of English and Bengali sentences) of E2B MT

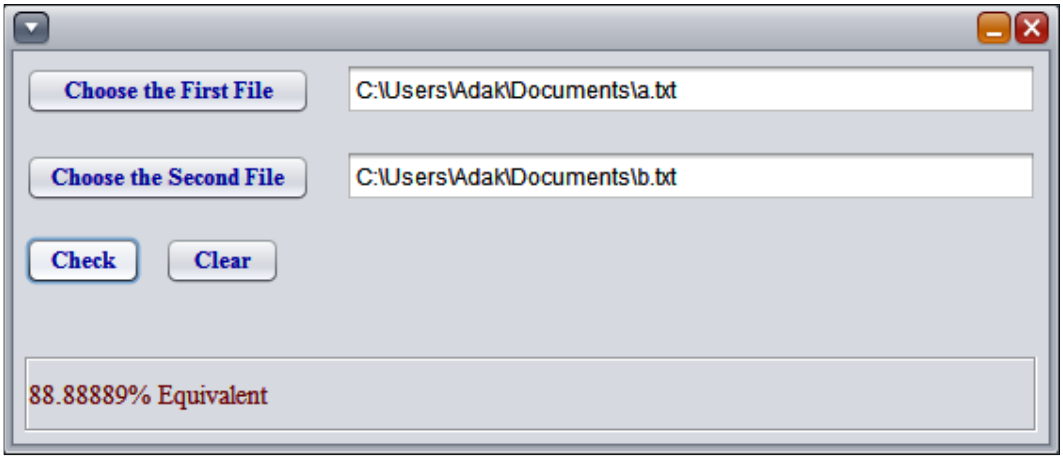

Fig.4(e) : Extended feature of E2B MT : file equivalency checking

4. ACCURACY MEASUREMENT

To test the accuracy our proposed E2B MT, we are taking help of *F-measure* technique.*Precision(Pr) = tp / (tp+fp)* , *Recall(Re) = tp / (tp+fn)* , Where *tp*= true positive, *fp*= false positive, *fn*=false negative.

$$F\text{-}Score = (2*Pr*Re) / (Pr+Re)$$

Table 1. F-Score Measurement

Measurement	*Precision(Pr)*	*Recall(Re)*	*F-Score*
M1	0.78	0.91	0.84
M2	0.73	0.87	0.79
Avg=(M1+M2)/2	0.755	0.89	**0.815**

So, our E2B machine translator has approx **81.5%** *F-Score Measure* accuracy.

5. CONCLUSIONS AND FUTURE WORKS

The proposed E2B-MT produces fairly good results for different English to Bengali sentence translations and the F-Score is more than eighty-one percent. But the main problem of this system is limited corpus. There are only near about two thousand words has been stored in the corpus. For this type of machine translation, we need all possible word stock. The huge amount of corpus gives the scope of good testing and then accuracy will be increasing. So our next venture of the E2B machine translation is to make the corpus more powerful and increasing the accuracy; And make the system genetic algorithm[14] and neural network[15] base, so the global optimality can be reached, and achieved a fast machine translator.

ACKNOWLEDGEMENTS

I would like to heartily thank Ms. SoumiChattopadhyay, Indian Statistical Institute, Kolkata-700108, India, for discussion various aspects of this paper.

REFERENCES

[1] P.F. Brown, S. A. Della Pietra, V.J. Della Pietra and R.L. Mercer ,*The Mathematics of Statistical Machine Translation: Parameter Estimation*, 19(2):263-311,1993.

[2] P.F. Brown, J. Cocke, S. A. Della Pietra, V.J. Della PietraF.Jelinek, J.C.Lai and R.L. Mercer ,*Method and System for Natural Language Translation*,1995.

[3] P.Koehn, *Pharaoh : A Beam Search Decoder for Phrase Base Statistical Machine Translational Models*, Proc. AMTA,2004.

[4] G. Doddington, *Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics.*Proc. 2nd Int. Conf.On Human Language Technology Research, 2002.

[5] K.Knight and J. Graehl. *Machine Transliteration* ,Computational Linguistics, 24(4):599-612, 1997.

[6] J.Wiebe, T.Wilson, R.Bruce, M.Bell, and M.Martin, *Learning subjective language*, Computational Linguistics, vol. 30, pp. 277–308 (2004).

[7] D. Jurafsky, J. H. Martin, *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson, ISBN : 978-81-317-1672-4

[8] SajibDasgupta, Abu Wasif, S. Azam, *An Optimal Way Towards Machine Translation from English to Bengali*, Proc. 7th Int. Conf. On Comp.And Inf. Tech. (ICCIT), 2004.

[9] F.J. Och. , *An Efficient Method for Determining Bilingual Word Classes,* In Proc. European Chap. of the Association for Computational Linguistics (EACL), 1999.

[10] Brill, E. (1997). *Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging*, In: Natural Language Processing Using Very Large Corpora, Kluwer Academic Press.

[11] Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. In: Proceeding of the Third Conference onApplied Natural Language Processing, ACL, Trento, Italy.

[12] Bill MacCartney, *NLP Lunch Tutorial: Smoothing*, 21 April 2005,
Website :http://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf .

[13] MuntsaPadr´o and Llu´ısPadr´o ,*Developing Competitive HMM PoS Taggers Using Small Training Corpora,* TALP Research Center, UniversitatPolit`ecnica de Catalunya, February 2004, Website : http://nlp.lsi.upc.edu/papers/padro04b.pdf .

[14] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc. , ISBN: 0-201-15767-5 .

[15] J.S. Jang, C.T. Sun, E. Mizutani, *Neuro-Fuzzy and Soft Computing, A Computational Approach to Learning and Machine Intelligence*, PHI, ISBN : 978-81-203-2243-1 .

AUTHOR

Mr. Chandranath Adak, Student Member, IEEE, is with the Department of Computer Science and Engineering, University of Kalyani, West Bengal- 741235, India.