Evaluation of Machine Translation Approaches to Translate English to Bengali

Shamsun Nahar
Dept. of CSE
World University of Bangladesh
Dhaka, Bangladesh
shamsun_nahar@ymail.com

Mohammad Nurul Huda
Dept. of CSE
United International University
Dhaka, Bangladesh
mnh@cse.uiu.ac.bd

Md. Nur-E-Arefin, Mohammad Mahbubur Rahman
Dept. of IT
IIT, University of Dhaka
Dhaka, Bangladesh
sami.arefin@gmail.com

***Abstract*—This paper describes the different types of machine translation (MT) approaches, where MT refers to the use of computers for the task of translating automatically from one language to another. It is highly challenging to build up a proper MT system which will works with full accuracy for translating foreign languages to native languages but this paper aims at providing a solution that could be helpful for building a MT system which will convert the English sentences into Bengali. Moreover, total 12 tenses such as- present indefinite, continuous, perfect, perfect continuous; past indefinite, continuous, perfect, perfect continuous; future indefinite, continuous, perfect and perfect continuous are used for the purpose of translating English sentence into Bengali that will require finding out the meaning from our own database. After comparing the experimental results based on different machine translation approaches with Google translator, it is found that one of our investigated as well as implemented methods, Corpus approach, provides higher accuracy in comparison with Google translator and other implemented methods.**

Keywords—Machine Translation; Machine learning; Natural Language Processing; Language Translation

I. INTRODUCTION

Bengali also known as Bangla is the mother tongue of Bangladesh. More than 220 million people speak in Bengali and it is ranked 7th most spoken language in the whole world. Bengali is also used in eastern area of India (West Bengal and Kolkata) as the medium of speaking and writing.

Numerous researches have done in the area of language translation but Natural Language Processing (NLP) is a quite tough job because of fully successful language translation machine. Natural languages are highly complex, mentioning that, words may have different meanings along with various use and translations, sentences may have distinct readings, and ambiguous relationships among linguistic entities. Since, it is a Human Language Technology (HLT) thus there are enormous prospects for doing research in this field. In fact, it is impossible to study on the whole language translation process at a time. As a result it requires to be segmented into many parts. Moreover, there is also another dilemma that most of them select a part of the source language for translating to the target language. In this paper English has been used as the source language and Bengali has been used as the target language because there are different types of sentences in both of these languages but the main focus of this paper is to evaluate some machine translation approaches by implementing them.

So far, a very few researches have done on English to Bengali language translation both in Bangladesh and West Bengal of India. Only the present indefinite and present continuous forms of English sentences are concerned in [1]. They represent a simple algorithm for language translation. Only one paper considered all forms of tenses [2]. Using Artificial Intelligence (AI) a Natural Language Processing (NLP) algorithm is proposed in [3]. In [4], Cockey-Younger-Kasami (CYK) algorithm is used for language translation where they used normal parse tree than the Chomsky Normal Form (CNF) parse tree because of some problems during the transformation phase. Morphological analysis is done in [5] where morphemes means minimal unit of meaning of grammatical analysis. A phrasal Example Based Machine Translation (EBMT) is described in [6]. Adaptive rule based machine translation between English to Bengali is used in [7]. In this paper they concentrated on rules which they found by proper translation from English to Bengali. Comprehensive Roman (English) to Bengali transliteration is defined in [8]. They actually designed a phonetics lexicon based English-Bengali transliteration. Verb based machine translation (VBMT), a new approach of machine translation (MT) from English to Bangla is proposed in [9].

II. MACHINE TRANSLATION

Machine translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. There are various approaches to Machine Translation. i) Word-for-Word translation, ii) The direct approach, iii) Transfer approach, iv) Corpus-based approach, v) Interlingua approach and vi) Statistical Machine Translation (SMT).

A. Word-for-Word Translation

Use a machine readable bilingual dictionary to translate each word in a text. An Example is given in Table I.

TABLE I. DICTIONARY

English	Bengali
I	আমি
Eat	খাই
Rice	ভাত

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

The advantages of this approach are Easy to implement, results gives a rough idea about what the text is about and the disadvantages are problems with word order means that this result in low quality translation rent designations.

B. Transfer Approach

The transfer model involves three stages a) Analysis, b) Transfer and c) Generation. In analysis stage the source language sentence is parsed and the sentence structure and the constituents of the sentence are identified. Example: I eat rice. Here words are: I, eat, rice and the sentence structure: [subject] [verb] [object]. In transfer stage transformation are applied to the source language parse tree to convert the structure to that of the target language (Fig. 1). Although there is some kind of 'transfer' in any translation system, the term **transfer method** applies to those which have bilingual modules between intermediate representations of each of the two languages [10].

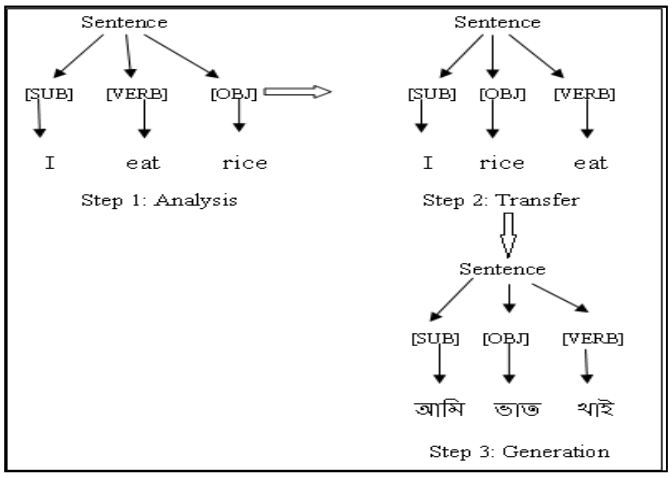

Fig. 1. Three stages of Transfer Approach

C. Direct Approach

The most primitive strategy is called the direct MT strategy, which is always between pairs of languages and based on good glossaries and morphological analysis. The direct approach lacks any kinds of intermediate stages in translation processes: the processing of the source language input text leads 'directly' to the desired target language output text [10]. Direct Approach has five steps to translate.

Example Sentence: You are playing football.

1. **Morphological analysis:** You playing Present Continuous football
2. **Identify constituents:** <You> <playing Present Continuous > <football>
3. **Reorder according to target language:** <You> <football> <playing Present Continuous >
4. **Look-up in the source target language dictionary:** <তুমি> <ফুটবল> <খেলছ>
5. **Inflect:** তুমি ফুটবল খেলছ

D. Corpus-based Approach

In corpus based MT (CBMT) approach two parallel corpora are available in source language (SL) and target language (TL) where sentences are aligned. First it is done by matching fragments against the parallel corpus and then adopting the method to the TL. Finally reassembling these translated fragments appropriately and then translation principle are applied. Fig. 2 shows an example.

Corpus-based Approach entails three steps:

1. Matching fragments against the parallel training corpora.
2. Adapting the matched fragments to the target language.
3. Recombine these translated fragments appropriately.

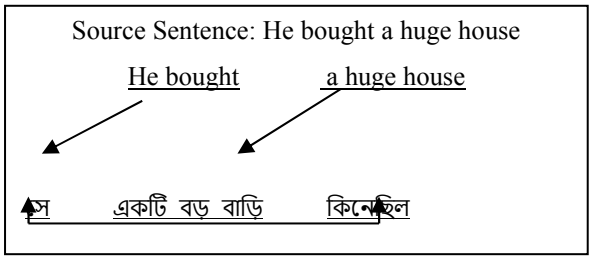

Fig. 2. Corpus-Based Approach

E. Interlingua Approach

The most advanced system is called the Interlingua MT strategy. In Interlingua method, the source text is analyzed in a representation from which the target text is directly generated. The intermediate representation includes all information necessary for the generation of the target text without 'looking back' to the original text [10]. The idea behind this approach is to create an artificial language, known as the Interlingua, which shares all the features and makes all the distinctions of all languages. To translate between two different languages, an analyzer is used to put the source language into the Interlingua, and a generator converts the Interlingua into the target language.

Two stages to follow for Interlingua Approach:

1. Extracting the meaning of a source language sentence in a language-independent form.

2. Generating a target language sentence from the meaning.

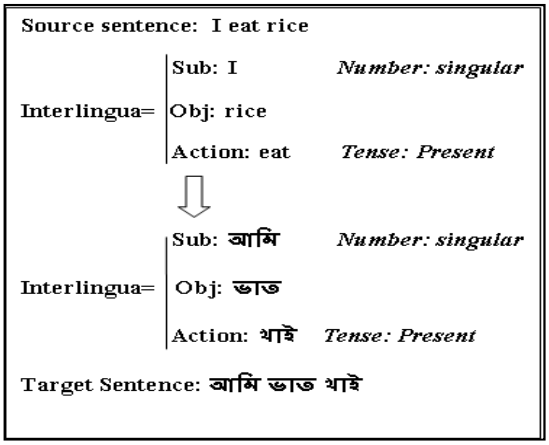

Fig. 3. Interlingua Approach

F. Statistical Machine Translation (SMT)

SMT models take the view that every sentence in the target language (TL) is a translation of source language (SL) sentence with some probability. SMT systems also deduce language and translation models from very large quantities of monolingual and bilingual data using a range of theoretical approaches to probability distribution and estimation [11]. The best translation of sentence is that which has the highest probability. In SMT three major components are, language model, translation model, search algorithm. If t-target language and s-source language then we can write,

$$P(t/s) = p(s/t)P(t)/P(s) \quad (1)$$

where $p(t/s)$ depends on the $P(t)$ which is probability of the kind of sentences that are likely to be in the language t. This is known as the language model $P(t)$. The way sentences in s get converted to the sentences t is called translation model $p(t/s)$.

III. IMPLEMENTED METHODS

Three (3) implemented methods of machine translation approach-i) Direct approach, ii) Corpus approach and iii) Transfer approach will be discussed here. These implemented methods can deal with multiline inputs and not case sensitive. Besides, all 12 tenses were taken into consideration while implementing these methods.

Different sentence pattern or structure can be dealt with these implemented methods. How sentence structure can change the meaning of a sentence is shown in the following Table II:

TABLE II. SENTENCE STRUCTURE

I play football	**আমি** ফুটবল খেলি
I have played football	আমি ফুটবল খেলেছি
I have a football	**আমার** ফুটবল আছে
You play football	**তুমি** ফুটবল খেলো
You have a football	**তোমার** একটি ফুটবল আছে
He play football	**সে** ফুটবল খেলে
He has a football	**তার** একটি ফুটবল আছে

From the table we can see that, if there is any verb after the subject or auxiliary verb then the meaning of "I" will be "আমি" but if there is no verb after auxiliary verb then the meaning of "I" will be "আমার". Same sentence structure is also applicable for 2nd person and 3rd person which are clearly seen from the table 2. These implemented systems can also deals with negative sentences and sentences with more than one object.

A. Direct approach

In direct approach the output for the sentence 'I am playing football in the field' is:

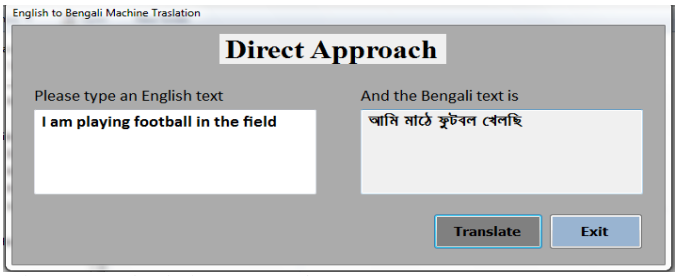

Fig. 4. Example of Direct Approach

The advantages of direct approach we get output more accurately (small amount of data set). But in case of huge data set this approach cannot achieve best result.

B. Corpus-based approach

The main advantage of this method is finding out the senses of words and phrases in different contexts in a speedy way. Moreover language users will find it very profitable as a corpus can provide them with a large collection of grammatical patterns, collocation and colligation of words and phrases to aid their analysis in a very short time. However the disadvantages are that corpora help in language learning and analyzing. Using corpora may, however, be a time-consuming task. The collections of texts in corpora may cause problems in analyses.

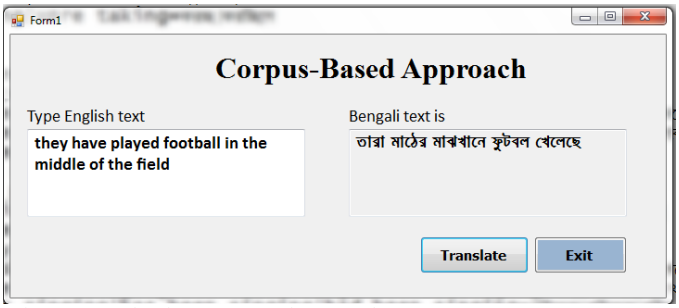

Fig. 5. Example of Corpus-Based Approach

C. Transfer Approach

It is possible with this translation strategy to obtain fairly high quality translations. However, sometimes it is not possible to show all the word meaning properly or sometimes in this approach some words are missing in the output text.

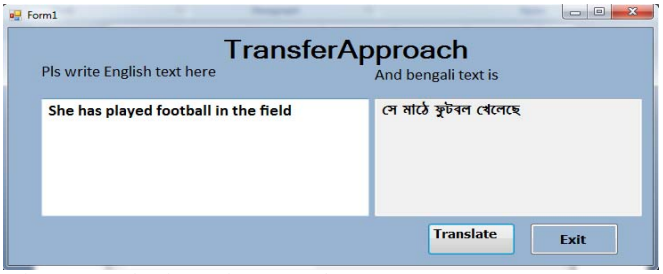

Fig. 6. Example of Transfer Approach

D. Tense based translation with implemented methods

In Table III, total twelve different sentences are translated from English to Bengali by our implemented methods. Each sentence represents a tense. Besides comparison with Google translator is also shown. Here bold indicates a wrong translation. So far our implemented systems (Direct, Transfer and Corpus) give best result with all 12 tenses in compare to Google translator.

TABLE III. TRANSLATION OF 12 TENSES FROM ENGLISH TO BENGALI AND COMPARE WITH GOOGLE TRANSLATOR

Name of Tense	English Sentence	Accurate Bengali Sentence	Direct Approach	Transfer Approach	Corpus Based Approach	Google Translator
Present	I play football	আমি ফুটবল খেলি	আমি ফুটবল খেলি	আমি ফুটবল খেলি	আমি ফুটবল খেলি	আমি ফুটবল খেলি
Present Continuous	I am playing football	আমি ফুটবল খেলছি	আমি ফুটবল খেলছি	আমি ফুটবল খেলছি	আমি ফুটবল খেলছি	আমি ফুটবল খেলছি
Present perfect	We have played Football	আমরা ফুটবল খেলেছি	আমরা ফুটবল খেলেছি	আমরা ফুটবল খেলেছি	আমরা ফুটবল খেলেছি	আমরা ফুটবল খেলেছি
Present perfect Continuous	We have been playing football for 2 hours	আমরা ২ ঘন্টা ধরে ফুটবল খেলতেছি	আমরা ২ ঘন্টা ধরে ফুটবল খেলতেছি	আমরা ২ ঘন্টা ধরে ফুটবল খেলতেছি	আমরা ২ ঘন্টা ধরে ফুটবল খেলতেছি	আমরা ২ ঘন্টা ধরে ফুটবল **খেলেছি**
Past	You played football	তুমি ফুটবল খেললে	তুমি ফুটবল খেললে	তুমি ফুটবল খেললে	তুমি ফুটবল খেললে	**আপনি ফুটবল খেলা**
Past Continuous	You were playing football in the field	তুমি মাঠে ফুটবল খেলছিলে	তুমি মাঠে ফুটবল খেলছিলে	তুমি মাঠে ফুটবল খেলছিলে	তুমি মাঠে ফুটবল খেলছিলে	**আপনি ক্ষেত্রের মধ্যে ফুটবল খেলা ছিল**
Past perfect	He had played football in the field	সে মাঠে ফুটবল খেলেছিল	**আমি** মাঠে ফুটবল **খেলেছিলাম**	**আমি** মাঠে ফুটবল **খেলেছিলাম**	সে মাঠে ফুটবল **খেলবে**	**তিনি ক্ষেত্রের মধ্যে ফুটবল খেলা ছিল**
Past perfect Continuous	He had been playing football in the field for 2 hours	সে মাঠে ২ ঘন্টা ধরে ফুটবল খেলিতেছিল	**আমি** মাঠে ২ ঘন্টা ধরে ফুটবল **খেলেছিলাম**	**আমি** মাঠে ২ ঘন্টা ধরে ফুটবল **খেলেছিলাম**	সে ২ ঘন্টা ধরে মাঠে ফুটবল খেলেতেছিল	**তিনি ২ ঘন্টা সময় ক্ষেত্রের ফুটবল খেলছিলেন**
Future	They will play football in the field	তারা মাঠে ফুটবল খেলবে	তারা মাঠে ফুটবল খেলবে	তারা মাঠে ফুটবল খেলবে	তারা মাঠে ফুটবল খেলবে	তারা **ক্ষেত্রের** ফুটবল খেলবে
Future Continuous	They will be playing football in the field for 3 hours	তারা ৩ ঘন্টা ধরে মাঠে ফুটবল খেলবে	তারা মাঠে ৩ ঘন্টা ধরে ফুটবল খেলবে	তারা মাঠে ৩ ঘন্টা ধরে ফুটবল খেলবে	তারা ৩ ঘন্টা ধরে মাঠে ফুটবল খেলবে	তারা তিন ঘন্টার **জন্য** মাঠে ফুটবল খেলবে
Future perfect	They will have played football	তারা ফুটবল খেলে থাকবে	তারা ফুটবল খেলে থাকবে	তারা ফুটবল খেলে থাকবে	তারা ফুটবল খেলে থাকবে	তারা ফুটবল **খেলেছে**
Future perfect Continuous	They will have been playing football for 3 hours	তারা ৩ ঘন্টা ধরে ফুটবল খেলতে থাকবে	তারা ৩ ঘন্টা ধরে ফুটবল খেলতে থাকবে	তারা ৩ ঘন্টা ধরে ফুটবল খেলতে থাকবে	তারা ৩ ঘন্টা ধরে ফুটবল খেলতে থাকবে	তারা ৩ ঘন্টার **জন্য** ফুটবল খেলবেন

Bold indicates wrong translation

IV. EXPERIMENTAL RESULT

The program which is used for finding the accuracy rate compares between two files; One is the original file and the other is implemented output file in different approaches like (direct, transfer, corpus). Initially the program counts sentence and word number in the original file. Comparison is basically done by word by word and sentence by sentence. Whether it finds any word mismatch then counts word mismatch and if it finds any sentence mismatch then counts sentence mismatch.

Finally the program counts word accuracy and sentence accuracy rate by the following equations 2 and 3,

$$\boldsymbol{Word\ accuracy\ rate} = \left(\frac{Total_{count_{word}} - Word_{mismatch}}{Total_{count_{word}}}\right) * \mathbf{100}\% \quad (2)$$

$$\boldsymbol{Sentence\ accuracy\ rate} = \left(\frac{Total_{count_{sentence}} - Sentence_{mismatch}}{Total_{count_{sentence}}}\right) * \mathbf{100}\% \quad (3)$$

Total 1027 sentences and total 5379 words are applied on three different machine translation approaches: Direct approach, Corpus Based approach, Transfer approach and Google translate. We get different accuracy rates from these approaches.

TABLE IV. ACCURACY RATE OF DIFFERENT APPROACH

	Direct Approach	Corpus Based Approach	Transfer Approach	Google Translate
Sentence Correct Rate (%)	68.4518	79.1626	77.9942	15.8715
Word Correct Rate (%)	80.8701	94.8689	91.4296	44.1718

From the Table IV, it is clear that sentence correct rate and word correct rate of Corpus approach is highest among all four methods and Google translate shows worst accuracy rate.

TABLE V. WORD AND SENTENCE COUNT OF DIFFERENT APPROACH

Different Method	Total word	Total sentence	Word mismatch	Line mismatch
Direct Approach	5379	1027	1029	324
Corpus Based Approach	5379	1027	276	214
Transfer Approach	5379	1027	461	226
Google Translate	5379	1027	3003	864

On Table V, we can see that total applied word is 5379 in Direct approach and word mismatch compare to original file is 1029. Moreover, total sentence is 1027 and sentence mismatch to original file is 324. The lowest number of word mismatch can be found from Corpus Based approach which is 276. Besides the lowest number of line mismatch can be found from the same approach which is Corpus based and the number is 214. So from this evidence it can be said that Corpus approach shows more accuracy among all those approaches.

A. Comparison with Related Work

With the extensive survey, we have noticed that there has not been much work carried out on different tenses. To be specific, there is little work done on different tenses. In Tense Based English to Bengali Translation Using MT System [2] paper considers total 12 different forms of tenses whereas this paper also worked with total 12 tenses. Moreover, that paper considered total 50*12=600 sentences while this work considered 1027 sentences. However, that paper's accuracy rate is less than this work. In paper [2], it has been observed that accuracy rates differ from tense to tense. For some tenses it is 100% but for some tenses it is 76%-90%. But our implemented work found 79.16% accuracy for all tenses by Corpus Based machine translation.

B. Why Corpus Based method is best?

There are two types of word files in this system. One is subject file and one is verb file. For each subject there is a flag correspond to the verb. For example, if the sentence is "I play" then the meaning of "play" is "খেলি", on the other hand if the sentence is "You play" then the meaning of "play" is "খেল". In Corpus based system it will take all the possible meaning of verb at first. Then most suitable meaning will be selected for the final translation. For example, the intermediate meaning of "I play football" in Corpus based method will be "আমি ফুটবল খেলি খেল". Then after final matching the final translation will be "আমি ফুটবল খেলি". That's why Corpus based method gives the best result.

V. CONCLUSION

This paper investigated some methods of machine translation and implemented them. One of our implemented methods, which is Corpus based method provides better result in comparison with Google translator and other implemented methods. The aim of this work is to find out that best method which is turn out to be Corpus, so that more work can be done on that particular method in future and make its accuracy rate higher. As it is a complicated work thereby it requires more improvement on detecting multiple Bengali meaning for an English word and on improving the artificial intelligence to detect phrases and idioms. Identifying multiple meanings, Phrase and idioms along with developing a strong data dictionary will be addressed in future. Working with interrogative sentences will also be taken into consideration in future.

REFERENCES

[1] S. Ahmed, M. O. Rahman, S. R. Pir, M. A. Mottalib, and Md. S. Islam, "A New Approach towards the Development of English to Bengali Machine Translation System," in International Conference on Computer Information and Technology (ICCIT) , pp. 360-364, Jahangirnagar University, Dhaka, Bangladesh, 2003.

[2] Kanija Muntarina, Md. Golam Moazzam and Md. Al-Amin Bhuiyan October (2013) "Tense Based English to Bangla Translation Using MT System" in International Journal of Engineering Science Invention ISSN (Online): 2319 –734, ISSN (Print): 2319 –6726 www.ijesi.org Volume 2 Issue 10 ǁ PP.30-38.

[3] S. A. Rahman, K. S. Mahmud, B. Roy, and K. M. A. Hasan, "English to Bengali Translation Using A New Natural Language Processing Algorithm," in International Conference on Computer Information and Technology (ICCIT), pp. 294-298, Jahangirnagar University, Dhaka, Bangladesh, 2003.

[4] S. Dasgupta, Abu Wasif, and S. Azam, "An Optimal Way of Machine Translation from English to Bengali" in ICCIT, 2004.

[5] A. N. K. Zaman, Md. A. Razzaque, and A. K. M. K. Ahsan Talukder, "Morphological Analysis for English to Bengali Machine Aided Translation" in National Conference on Computer Processing of Bangla, Dhaka, Bangladesh, 2004.

[6] S. K. Naskar, and S. Bandyopadhyay, "A Phrasal EBMT for Translation English to Bengali," in MT Summit X, Kolkata, India, 2005.

[7] Judith Francisca, Md. Mamun Mia, and Dr. S. M. Monzurur Rahman, "Adapting Rule Based Machine Translation From English to Bangla" in Indian Journal of Computer Science and Englineering (IJCSE), 2(3), pp.334-342, 2011.

[8] Naushad UzZaman, Arnab Zaheen, and Mumit Khan, "A comprehensive Roman (English)-to-Bangla transliteration scheme." 2006.

[9] Masud Rabbani, Kazi Md Rokibul Alam, and Muzahidul Islam. "A new verb based approach for English to Bangla machine translation." In Informatics, Electronics & Vision (ICIEV), International Conference on, pp. 1-6. IEEE, 2014.

[10] William John Hutchins, and Harold L. Somers, "An introduction to machine translation" in London: Academic Press, Vol. 362, 1992.

[11] Andy Way, and Nano Gough, "Comparing Example-Based and Statistical Machine Translation" in Natural Language Engineering, Vol. 11(3),pp.295-309,2005