Word Sense Based Hindi-Tamil Statistical Machine Translation

Article *in* International Journal of Intelligent Information Technologies · October 2018

DOI: 10.4018/IJIIT.2018010102

2 authors:

Vimal Kumar K
University of Limerick
14 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Divakar Yadav
Indira Gandhi National Open University (IGNOU)
156 PUBLICATIONS **1,574** CITATIONS
SEE PROFILE

Word Sense Based Hindi-Tamil Statistical Machine Translation

Vimal Kumar. K*, Divakar Yadav
Department of CSE & IT, Jaypee Institute of Information Technology, India

ABSTRACT

Corpus based natural language processing has emerged with great success in recent years. It is not only used for languages like English, French, Spanish, and Hindi but also is widely used for languages like Tamil, Telugu etc. This paper focuses to increase the accuracy of machine translation from Hindi to Tamil by considering the word's sense as well as its part-of-speech. This system works on word by word translation from Hindi to Tamil language which makes use of additional information such as the preceding words, the current word's part of speech and the word's sense itself. For such a translation system the frequency of words occurring in the corpus, the tagging of the input words and the probability of the preceding word of the tagged words are required. Wordnet is used to identify various synonym for the words specified in the source language. Among these words, the one which is more relevant to the word specified in source language is considered for the translation to target language. The introduction of the additional information such as part-of-speech tag, preceding word information and semantic analysis has greatly improved the accuracy of the system.

Keywords **–** Machine Translation, Naïve Bayes method, Statistical approach, Word Sense Disambiguation, Latent Semantic Analysis

INTRODUCTION

Natural language processing (NLP) is a part of artificial intelligence which interacts with the systems (computer) through natural languages to perform desired actions. It deals with understanding and analyzing human languages in order to perform various functionalities which can enhance the interaction between the machine and the humans. There are various widely used algorithms under NLP, especially statistical natural language processing but each algorithm has its own bottleneck. These algorithms are usually based upon the analysis of large textual corpora and then calculating probabilities in order to achieve the desired results. According to linguistics, corpus refers to large structured texts consisting of numerous words which are used for statistical analysis of the text. Generally, the corpus should be annotated to provide an efficient statistical analysis. The Corpus consists of each and every word in every sentence used for the language analysis. These words are added to the corpus along with the information about its part of speech such as: verbs, adjectives, nouns, and adverbs etc., which are called as POS tags. Corpus based NLP techniques have emerged with great success in the recent years. It is not only used for languages like English, French, Spanish, and Hindi but also is widely used for languages like Tamil, Vietnamese etc.

Numerous algorithms have been introduced in statistical machine translation to provide various intelligent functionalities for human-computer interaction. All these algorithms parse the sentences and then group the words before the translation process. Parsing of free word order languages such as Indian is also a bottleneck in these methods (A. Bharati et. al, 2009; A. Bharati & R Sangal, 1993). Local word grouping (LWG) is basically used in Indian languages since there is a need for grouping the words based on the context in which it is used and the meaning of those words will be clear only when it is grouped together (A. Bharati et. al, 1991; P R Ray et. al, 2003). In these existing technologies related to corpus based NLP, the statistical analysis for machine translation makes use of a parallel corpus. In a parallel corpus, each word in the source language is mapped parallel with its corresponding word in the target language. In addition to parallel corpus being used for translation, the part-of-speech of the words is also considered for machine translation. Also, in statistical machine translation, the target texts are generated on the basis of statistical models and these models are derived from the analysis of the text corpus of the two languages. Generally, a document is translated to a probable sentence in the target language according to the probability distribution $P(t|h)$ which refers to the probability of string t in the target language (for example, Tamil) given the string h in the source language (for example, Hindi).
Naive Bayes algorithm is one of the existing algorithms which are based on statistical analysis of the existing

bilingual corpus. The algorithm uses probability of occurrence of words for translation from one language to another. For a particular word, its probability is calculated based upon the frequency of occurrence of the word in the corpus. The meaning which has maximum occurrence in the target language will be the probable translation for the input word. The mathematical representation of this algorithm is:

$$\text{Posterior} = \text{Likelihood} * \text{Prior Evidence}$$

$$P(\frac{x}{y}) = P(\frac{y}{x}) * P(x) * P(y)$$

Since P(y) will not affect the result, the equation is equated as shown below,

$$P(\frac{x}{y}) \cong P(\frac{y}{x}) * P(x)$$

This mathematical representation signifies that there is translation model and language model being used to perform the translation process. Translation model analyzes the translation between source and target language whereas the language model analyzes the target language being used. These statistical algorithms can further be improved by considering the tag information of the word under consideration but still there will be lack of efficiency and accuracy. Also ambiguity is one of the concerns of these algorithms. In order to overcome this ambiguity issue, an analysis on these languages has been made and found that the sense of word can be used to eradicate this ambiguity issue. Certain word has different translations based on its part-of-speech which has been included in this Naïve Bayes model. There are words that have multiple translations based on the context in which it is used and this can be identified using word sense disambiguation over the sentence. Thus, the objective of this system is to improve the efficiency of the statistical machine translation by providing additional information such as the part-of-speech of the word to be translated as well as the information about the preceding words and also to make use of the word's sense in the translation process. Since this translation process takes in to account the word's part-of-speech (syntactic information) as well as the word's sense (semantic information) during the machine translation, it improves the efficiency and accuracy of the translation.

Rest of the contents of this paper is arranged as follows: Section 2 discusses the existing works, researches and technologies which have been prevalent in the field of machine translation. It also discusses about challenges faced and solutions proposed in the respective field. Section 3 consists of a detailed description of the algorithm introduced, theory, implementation and mathematical interpretation. The results and its analysis are described in section 4. Finally in section 5, it concludes the work along with some future directions.

RELATED WORK

Various techniques have been proposed in the sub field of inter language translation under the umbrella of natural language processing (NLP). Cross linkage of words from two different languages have empowered NLP facilitating bilingual dictionaries and other practical applications. This can be a challenging task as the linkage needs to be balanced and two languages can be varied on grounds of culture, meanings etc. Hindi to English linkage suffers from kinship relations (uncle in English can be mama or chacha in Hindi), musical instruments, grains, kitchen utensils etc. Direct and hypernamy linkage can cater to such challenges. Also manual linking of two languages can be facilitated using WordNet SynsetMatcher tool (Jaya Saraswati et. al, 2010).

Annotation in Hindi and English varies highly as both the languages differ in terms of grammar. The two techniques for tagging can be: linking all possible words in both source and target language, or, linking least possible words in both source and the target language. Since exact correspondence is hard to find and word to word translation is possible for very few words, thus remaining words are translated based on the combination of words. Thus

annotation can follow certain guidelines such as following fuzzy, regular and null links in punctuation which can lead to desired results (Rahul Kumar Yadav & Deepa Gupta, 2010).

A system, ANUBHARTI which follows an approach of machine aided translation having the combination of example based and corpus based approaches with some elementary grammatical analysis is been developed for corpus generation (A. Bharati et. al, 1997). In ANUBHARTI the traditional EBMT approach has been modified to reduce the requirement of a large example based corpus. ANUBHARTI-II in 2004 uses Hindi as a source language for translation to other Indian language.

A machine translation system which translates pure (standard) Hindi to pure English was implemented in 2004 by incorporating additional level to AnglaBharti-II and AnuBharti-II. The system gives satisfactory results in more than 90% cases (R. M. K. Sinha & Anil Thakur, 2005; R. M. K. Sinha, 2004).

Vishal Goyal and G.S. Lehal (2010) proposed a machine translator from Hindi to Punjabi. The methodology used for the translation was direct translation and later on improving the language learning modules for the enhancement of the quality of the system. The accuracy of the translation is approximately 95%. The web tool developed has many application areas like over internet for the particular community of peoples, for the newspapers and sending e-mails from Hindi to Punjabi or vice versa (Vishal Goyal, 2010; Vishal Goyal & G. S. Lehal, 2010).

The performance of the multilingual information retrieval employing dictionary based strategy for query translations are measured and analyzed in the paper authored by Raju Korra et. al, (2011). The experimental result shows that the performance of the multilingual information retrieval system enhances over the monolingual information retrieval system by 31.4%.

For English to Hindi statistical machine translation, the key challenge is that the Hindi language is richer in morphology than the English language. Reordering of English source sentences in accordance with Hindi language and/ or making use of suffixes of Hindi words are the two strategies that can facilitate reasonable performance. Since Indian languages and English differ in word order, morphologically rich Indian languages and unavailability of huge parallel corpora for two languages make the above two strategies more challenging to derive the desired results with reasonable performance (Ananthakrishnan, R. et. al, 2009).

PROPOSED METHOD

This proposed method performs translation at the word level and it follows a step by step procedure. The paper focuses on a system which makes use of a translation model and a language model. The translation model is developed to provide the translation from one language to another, based on the analysis of both the languages. Since the translation process depends on the grammatical sequence of words in the source language, an analysis of grammatical structure of the source language is performed and a language model is developed for the same. It is identified for the languages under consideration that the word's sense also matter during the translation process, the word's sense are identified using the semantic analysis. The system architecture is as shown in figure-1.

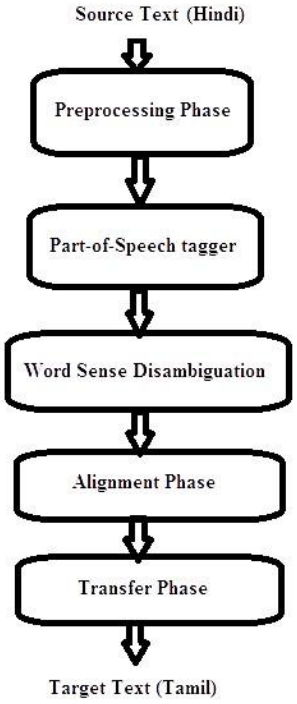

Figure 1 System Architecture

Following are the detailed description about the various phases of this proposed translation system,

Preliminary Phase - Corpus Generation

The text corpus is a large set of structured text electronically stored in the database. The corpus is used for the statistical analysis within the language territory. In this algorithm, a multi lingual database is built for the language model and translation model by traversing over a big textual corpus in both the languages provided by Technology Development for Indian Languages **(**TDIL**)** Programme, Department of Information Technology (DIT), MC&IT, Govt of India. The database is made in accordance to the part-of-speech tagging for individual words in the source language (Hindi) along with semantically equivalent of the target word (Tamil). Additionally words' frequency of occurrence in the text corpus is also stored. This is one part of corpus generation which is required for the translation model. Another database is built for Tamil text by traversing a Tamil corpus and storing values in the database for preceding words, current word, and their frequencies. This database signify the language model required for the statistical system

Pre-processing Phase

Since this system is a word level translation system, the preprocessing phase is to tokenize the input text in to individual sentences. These sentences are further tokenized to extract the words and these words are passed on to the next phase of the translation system which is the part-of-speech tagging phase

POS Tagging Phase

In machine translation, the part-of-speech tagging has to be performed that refers to the marking of the words for information based upon the context of the word and also the relationship between the word and the adjacent words. It refers to the identification of words as nouns, verbs, adjectives, adverbs etc. Part-of-speech tagging is generally of three types namely: rule based part-of-speech tagging, transformation based part-of-speech tagging and probabilistic based part-of-speech tagging. Ambiguity is one of the major problems that come across while performing part-of-speech tagging. This system uses the well known HMM based POS Tagging. The tagger considers the preceding word information to improve the tagging process that reduces the ambiguity. As the proposed system considers the tagging information during translation, the accuracy of the system depends more on this phase of the translation process. In case the tag assigned to a particular word is wrong, then the system's translation won't be proper.

Word Sense Disambiguation using Hindi Wordnet

This phase uses the Hindi Wordnet dictionary to identify the various senses for the particular word with its identified part-of-speech. Semantic analysis between the words under consideration over the various senses of the word provides a score of its relevance. The high scored senses are used further in the translation. For the sense disambiguation, the latent semantic analysis (LSA) technique is used. LSA technique is a singular value decomposition method which decomposes the term-frequency matrix in to three different matrices namely left singular matrix (L), right singular matrix (R) and singular diagonal matrix (D). Left singular matrix is a matrix which has the Eigen vectors of the term matrix and its transpose and this matrix is also called as term matrix. Right singular matrix is one which contains the singular values of the term frequency matrix as the diagonal elements and it is also called as concept matrix. Whereas the singular diagonal matrix has the Eigen vectors of the transpose of term-frequency matrix and this matrix is known as concept by document matrix.

$$\begin{bmatrix} f_{t1}^{1} & f_{t2}^{1} & \ldots\ldots & f_{tn}^{1} \\ f_{t1}^{2} & f_{t2}^{2} & \ldots\ldots & f_{tn}^{2} \\ \ldots & \ldots & \ldots & \ldots \\ f_{t1}^{m} & f_{t2}^{m} & \ldots\ldots & f_{tn}^{m} \end{bmatrix} = L * R * S$$

where, f_{tn}^{m} - indicates the frequency of n^{th}-term in m^{th}-sentence

The term frequency of words is arranged in the form of a matrix with column as different terms (words) used in the text and row of the matrix as the sentence identification numbers. After the decomposition of this matrix, the product of the right singular matrix and singular diagonal matrix gives the frequency of semantically similar words in different sentences and the similarity between two sentences is found using cosine similarity over the product of right singular vector and singular diagonal vector of the sentence considered (K. Vimal Kumar et. al, 2015). The cosine similarity is calculated using the below mentioned in the equation

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i * y_i}{\left(\sqrt{\sum_{i=1}^{n} x_i^2} * \sqrt{\sum_{i=1}^{n} y_i^2}\right)}$$

At this phase, the system has identified the relevant senses for the word, now the system has to rearrange the sentence since the naïve Bayes method requires the proper sequence in the target language.

Alignment Phase (Syntactic Restructuring)

Based on the analysis, it is found that Hindi language follows Subject Object Verb (SOV) sequence that is same as in Tamil. The transfer phase makes use of the target language model and considers the preceding word information in the target language level. Since, the text which is in source language follows the grammatical correctness of the words in target language; there is no need for rearranging the words based on the target language grammar under certain exceptional cases such as compound sentences, which gives emphasis on the restructuring of sentences in the target language Tamil. In Tamil language there are various affixes being used in a sentence which doesn't exist in Hindi language. Accordingly the grammatical restructuring of the words in source language happens such that the source language sentence is aligned with the grammar as in the target language. This is achieved by applying the statistical based alignment algorithm. Basically, this system makes use of GIZA++ (a tool for word alignment). This tool generates the alignment table which is required for the statistical method. The sample alignment table is as shown in table-1 which has the position in target language, source word position, length of Hindi sentence and length of its corresponding Tamil word.

Naïve Bayes model is applied over this alignment table to identify the probable alignment in target language for the given sentence in source language. The one which has the higher probability will be used for aligning these source language words.

TABLE I – Sample Alignment Table

S.No	J (Word's Position in target language)	I (Word's Position in Source language)	L (Source Sentence Length)	M (Target Sentence Length)	Frequency
1	4	3	8	10	252
2	7	7	10	10	231

Transfer Phase

The basic model of language translation, also called as parallel translation, finds every individual target language word of the source language word in a list and substitutes this word with the equivalent target language word. But this method does not ensure accurate results as the relationship followed between words in source language and in target language are often one-to-many, instead of one-to-one. This is due to that a word in source language can give different equivalent translations in the target language based on the context where it is being used. Solutions for increasing the accuracy of the results are to distinguish the word on the basis of its part-of-speech and also on the basis of word's sense. In this proposed system, the part-of-speech of the word is used to identify the category of words and in the particular category a word can have multiple senses which are further identified using LSA and Wordnet tool. The sense of the word is used in our statistical translation model.

Probability serves a major role to analyze any complex data. This system follows analysis for translation by taking it into account the probability of target word given the source word and its corresponding part-of-speech. The probability can be represented as mentioned below,

$$P(W_t \mid W_s, W_{tag}) = \arg\max_{j \to 0 to N} P(W_j, W_{tag} \mid W_t) * P(W_t)$$

Where, N indicates the total number of senses identified for word W_s

$$P(W_s, W_{tag} \mid W_t) = P(W_s \mid W_t) * P(W_{tag} \mid W_t)$$

$$P(W_s \mid W_t) = \frac{count(W_s, W_t)}{count(W_t)}$$

$$P(W_{tag} \mid W_t) = \frac{count(W_{tag}, W_t)}{count(W_t)}$$

$$P(W_t) = P(W_t \mid W_{prev\_word})$$

$$P(W_t) = \frac{count(W_t, W_{prev\_word})}{count(W_t)}$$

The above mentioned formulae are used for finding the probable translation for one sense of the word in the source language. To find the most probable translation for that context in which it is being used, the above equation is applied over all the various senses of the word and the one which has the maximum probability value is selected as the best translation for the word. Sample tagged transfer database is as shown in table-2. This database has Hindi words, its corresponding Tamil words based on the POS tag of the Hindi word along with the frequency of occurrence. In table-2, the word अच्छी (*AACHCHI*) appears in 89 sentences out of which it is converted to நல்ல (*NALLA*) in 49 sentences and the rest of the sentences it is converted as நன்றாக (*NANDRAGA*). From this it is evident that the words' part of speech can be used to distinguish the target word translation. The same word appearing as adjective, has two different equivalent words in Tamil in *60* cases and in such cases it can be narrowed to one particular word by considering its preceding word information as well as the word's sense.

TABLE II Sample Transfer Database

Hindi word	Tamil word	Tag	Frequency
अच्छा	நல்ல	JJ	53
अच्छी	நன்றாக	RB	29
अच्छी	நன்றாக	JJ	11
अच्छी	நல்ல	JJ	49
उत्पादक	உற்பத்தியாளர்	N_NN	24

TABLE III Sample Precedence Database

Precedence word	word	frequency
^	நல்ல	18

மிகவும்	நல்ல	25
^	நன்றாக	13
இது	நன்றாக	15
^	உற்பத்தியாளர்	4
நல்ல	உற்பத்தியாளர்	12

There comes the need for a precedence database whose sample is as shown in table-3 and this database has the details about the preceding word for any given Tamil word along with its frequency. The use of ^ symbol for precedence word denotes that the word occurs in the beginning of any sentence.

For example,
Consider the input text, **"ताजा साँसें और चमचमाते दाँत आपके व्यक्तित्व को निखारते हैं।"**
After preprocessing the text, individual words are identified as unique tokens from the input text. This tokenized text is subjected to POS Tagging. The tagged sentence output is ताजा\JJ साँसें\N_NN और\CC_CCD चमचमाते\JJ दाँत\N_NN आपके\PR_PRP व्यक्तित्व\N_NN को\PSP निखारते\V_VM हैं\V_VAUX ।\RD_PUNC. This tagged text is considered for syntactic restructuring but in this case there won't be any change in the order of the sentence so the same output is fed to statistical transfer phase and is described as follows,
Considering the word *"ताजा\JJ"*,

From the wordnet the various senses that are found for (ताजा as Adjective) are - ताज़ा, ताजा, अम्लान, अशुष्क, आला, गरमागरम, टटका. These words are further subjected to word sense disambiguation using LSA. For the latent semantic analysis, the sentences formed using these various senses and the given word is converted in to term-frequency matrix which is used in the latent semantic analysis. The LSA algorithm identifies that ताजा and ताज़ा are the relevant words for the given input word.

Case-1: When source word W_s is "ताजा" and Tag W_{tag} is "JJ" and target word W_t is "புத்துணர்ச்சியான" (Puthunarchiyana means Fresh), then $P(W_s, W_{tag} \mid W_t) = 30/34$
Also when the target word W_t is "புத்துணர்ச்சியான" and precedence is nothing, then $P(W_t) = 22/30$
So combined probability $P(W_t \mid W_s, W_{tag}) = (30/34)*(22/30) = 0.647$

Case-2: When source word W_s is "ताज़ा" and Tag W_{tag} is "JJ" and target word W_t is "**புதுப்பித்து**" (Puthupithu means New), then $P(W_s, W_{tag} \mid W_t) = 8/10$
If the target word W_t is **புதுப்பித்து** and precedence is nothing, then $P(W_t) = 1/10$
So combined probability $P(W_t \mid W_s, W_{tag}) = (8/10)*(1/10) = 0.08$

Out of these probabilities first one is highest so the target text is "புத்துணர்ச்சியான"

Similarly, the other words are translated. The final translated output will be புத்துணர்ச்சியான\JJ சுவாசம்\N_NN மற்றும்\CC_CSD பளபளப்பான\JJ

பற்கள்\N_NN தங்களின்\PR_PRP தோற்றத்தை\N_NN நிர்ணயிக்கிறது\V_VM_VF .\RD_PUNC

RESULTS & EVALUATION

This system has been evaluated using various sentences (restricted to health domain) and the Bilingual Evaluation Understudy (BLEU) score is calculated to be 0.68. The precision and recall of the system are found to be 0.85 and 0.7 respectively. As shown in the table-IV, the system is found to give very good results by introducing word sense based translation which is compared with its counterpart without word sense disambiguation (WSD). The precision and recall gradually improves with increase in corpus size. This is due to inclusion of additional words to the corpus which makes it mature enough compared to the lesser corpus size.

TABLE IV Comparison of the Precision & Recall

S. No.	Corpus Size (in number of words)	Statistical Machine Translation (with WSD)		Statistical Machine Translation (without WSD)	
		Precision (in %)	Recall (in %)	Precision (in %)	Recall (in %)
1	10000	65	63	58	57
2	20000	76	72.5	71	72
3	30000	87	86.5	83	79

CONCLUSION AND FUTURE WORK

The basic advantage is that the algorithm considers the sense of the word during translation process. In addition to that, the introduced algorithm also tends to increase accuracy of the translation as compared to parallel translation or any other existing methods for language translation. The algorithm calculates probability of occurrence for individual word, precedence word and tagging of words during the translation. The precision and recall of the algorithm are 87% and 86.5% respectively, which shows considerable improvement compared to the system without word sense disambiguation, whose precision and recall are found to be 83% and 79% respectively.

Although the algorithm tend to improve the results, but still there is scope for modifications that can further improve the accuracy of the output. One of such modifications can be introducing good part-of-speech tagging in the source language. Another modification can be making the algorithm machine intelligent so that if the input word is not contained in the corpus, the algorithm intelligently derives proper translation. These further modifications, if implemented in future can improve results in machine translation from Hindi to Tamil.

REFERENCES

[1] Bharati, A., Husain, S., Misra, D., & Sangal, R. (2009). Two stage constraint based hybrid approach to free word order language dependency parsing. *Proceedings of the*

11th International Conference on Parsing Technologies - IWPT '09. doi:10.3115/1697236.1697251.

[2] Bharati, A., & Sangal, R. (1993). Parsing free word order languages in the Paninian framework. *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* -. doi:10.3115/981574.981589

[3] A. Bharati, Chaitanya Vineet, P. Amba Kulkarni, Rajeev Sangal (1997) 'Anusaaraka,"Machine translation in Stages"', *A Quarterly in Artificial Intelligence*, 10(3), pp. 22-25.

[4] Ananthakrishnan, R., Bhattacharyya, P., Hegde, J. J., Shah, R. M., and Sasikumar, M. (2008) 'Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation', *Proceedings of International Joint Conference on Natural Language Processing*.

[5] A Bharati, Vineet Chaitanya, and Rajeev Sangal (1991) Local Word Grouping and Its Relevance to Indian Languages, *In Frontiers in Knowledge Based Computing (KBCS90)* V. P. Bhatkar and K. M. Rege edn., New Delhi: Narosa Publishing House.

[6] Hemant Darbari (1999) 'Computer-assisted translation system – an Indian perspective', *Machine Translation Summit VII*, Kent Ridge Digital Labs, Singapore, pp. 80-85.

[7] Jaya Saraswati, Rajita Shukla, Ripple P. Goyal, Pushpak Bhattacharyya (2010) 'Hindi to English Wordnet Linkage: Challenges and Solutions', *8th International Conference on Natural Language Processing*.

[8] Kumar, K. V., Yadav, D., & Sharma, A. (2015). Graph Based Technique for Hindi Text Summarization. *Advances in Intelligent Systems and Computing Information Systems Design and Intelligent Applications*, 301-310. doi:10.1007/978-81-322-2250-7_29

[9] P. R Ray, V. Harish, A. Basu, S. Sarkar (2003) 'Part of speech tagging and local word grouping techniques for natural language parsing in Hindi', *International Conference on Natural Language Processing*.

[10] Kumar, P., & Goyal, V. (2010). Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments. *International Journal of Computer Applications*, 5(9), 15-19. doi:10.5120/941-1319

[11] R. M. K. Sinha, Anil Thakur (2005) 'Machine Translation of bi-lingual Hindi-English (Hinglish) text', *In proceedings of the tenth Machine Translation Summit*, MT Summit X, Phuket, Thailand, pp. 149-156.

[12] R. M. K. Sinha (2004) 'An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II architectures', *In proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS)*, pp. 1-9.

[13] Yadav, R. K., & Gupta, D. (2010). Annotation Guidelines for Hindi-English Word Alignment. *2010 International Conference on Asian Language Processing*. doi:10.1109/ialp.2010.58

[14] Korra, R., Sujatha, P., Chetana, S., & Kumar, M. N. (2011). Performance evaluation of Multilingual Information Retrieval (MLIR) system over Information Retrieval (IR) system. 2011 *International Conference on Recent Trends in Information Technology (ICRTIT)*. doi:10.1109/icrtit.2011.5972453

[15] Pandian, S. L., & Geetha, T. (2008). Morpheme based Language Model for Part-of-Speech Tagging. *Polibits, 38*, 19-25. doi:10.17562/pb-38-2

[16] Saravanan, K., Parthasarathi, R., Geetha, T.V. (2003) 'Syntactic Parser for Tamil', *Proceedings of the Tamil Internet Conference*, Chennai, Tamilnadu, India, pp. 28-37.

[17] V. Goyal and G. S. Lehal (2009) 'A Machine Transliteration System for Machine Translation system : An Application on Hindi-Punjabi Language Pair', *Atti Della Fondazion Giorgio Ronchi (Italy)*, LXIV(1), pp. 27-35.

[18] Goyal, V., & Lehal, G. S. (2008). Hindi Morphological Analyzer and Generator. 2008 *First International Conference on Emerging Trends in Engineering and Technology*. doi:10.1109/icetet.2008.11

[19] Vishal Goyal (Oct 2010) Development of a Hindi to Punjabi Machine Translation System. [Online]. Available at: http://www.languageinindia.com.

[20] Goyal, V., & Lehal, G. S. (2010). Web Based Hindi to Punjabi Machine Translation System. *Journal of Emerging Technologies in Web Intelligence JETWI*, 2(2). doi:10.4304/jetwi.2.2.148-151.