Word Sense Based Approach for Hindi to Tamil Machine Translation Using English as Pivot Language

Article *in* International Journal of Advanced Intelligence Paradigms · January 2018

DOI: 10.1504/IJAIP.2018.10008778

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 approach for Hindi to Tamil machine translation using English as pivot language

K. Vimal Kumar* and Divakar Yadav

Department of Computer Science Engineering,
Jaypee Institute of Information Technology,
Noida-201 307, Uttar Pradesh, India
Email: vimalkumar.k@gmail.com
Email: divakar.yadav0@gmail.com
*Corresponding author

Abstract: Machine translation is defined as the translation of source text to a desired target text. As there is resource availability in different languages in the internet world, there is need to share the knowledge to a different set of audience who knows only their native language. Proposed system is aimed to build a word sense-based statistical machine translation system (Hindi to Tamil). Since there is a lack of resources in these languages, there is need of some other intermediate pivot language which has high resource availability and English language is chosen as one. Initially, the Hindi text is subjected to pre-processing phase where the text is morphologically and syntactically analysed. After analysis, the senses of the words are identified using latent semantic analysis (LSA) in order to provide a meaningful translation. Once these analysis are done, the sentence is subjected to statistical translation from source to target language through the intermediate pivot language. This system has an improved efficiency when compared with the system that does not have sense identification and pivot language.

Keywords: statistical machine translation; word sense disambiguation; latent semantic analysis; LSA; pivot-based machine translation.

Reference to this paper should be made as follows: Kumar, K.V. and Yadav, D. (xxxx) ‘Word sense-based approach for Hindi to Tamil machine translation using English as pivot language’, *Int. J. Advanced Intelligence Paradigms*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Vimal Kumar has completed his Masters of Engineering from College of Engineering, Guindy, Chennai. Currently, he is working as an Assistant Professor in Jaypee Institute of Information Technology, India. He has more than eight years of experience in teaching and research work. His areas of interest are natural language processing, machine translation and AI.

Divakar Yadav received his PhD in Computer Science and Engineering from Jaypee Institute of Information Technology, Noida, India in Feb 2010. He spent one year, from Oct 2011 to Oct 2012, in Carlos III University, Leganes, Madrid, Spain as a Post Doctoral Fellow. He has published more than 60 research papers in reputed international/national journals and conference proceedings. His areas of interest are information retrieval and soft computing.

1 Introduction

Machine translation is a process which encodes the target language text from the source language text without losing the semantics mentioned in the source text. Since there are around 7,000 languages in this world, the need for machine translation over any combination of languages has greater necessity. The requirement for the machine translation system has increased due to need for technical communication across the globe in the current generations. The machine translation system basically requires certain manually translated sentence as a reference between the languages under consideration. Based on these manually translated sentences, the new sentence can be translated using various approaches of machine translation. The various approaches of machine translation are: rule-based machine translation, statistical machine translation, transfer-based machine translation and hybrid machine translation. In case of rule-based machine translation, a corpus is required for rule generation phase. Based on the generated rules, the system translates the text from one language to another. In a statistical machine translation, the system uses a parallel corpus to generate the statistical information about various words in both languages. Transfer-based machine translation makes use of an intermediate language between the source and target language. So, it requires a corpus that has the relationships between these three languages, under consideration. The issue with these kinds of corpus is its unavailability in electronic form for corpus generation in different languages. This issue can be overcome by introducing a pivot language in between the source and target language and the pivot language should be one which has rich resource available in electronic form. In this paper, English language has been identified as a pivot language because of its rich resource availability and also it can be mapped with the Indian languages – Hindi and Tamil.

The mapping between Indian languages and the pivot language (English) is also a bottleneck in this research as the grammar for these languages are totally different. Moreover, English is a fixed word order language whereas Indian languages (Hindi and Tamil) are free word order languages. To overcome this issue, the word alignments between the languages under consideration are required. There is an existing word alignment algorithm known as IBM model on word alignment (Brown et al., 1993). These IBM models never include the part-of-speech of the word to find the alignment parameters. Based on the analysis, it is found that the words alignment parameters changes with respect to its part-of-speech. Thus in this proposed system, we introduce a modified word alignment algorithm which considers the part-of-speech of the word too. Once the words are aligned, the proposed system performs a transfer from Hindi to English language using naïve Bayes method. But the performance of naïve Bayes method degrades the systems accuracy as there is more semantic distortion because of the pivot language, which can be overcome by the introduction of semantic analysis to interpret the words sense in the input language. The same naïve Bayes method is used from the transfer from English to Tamil as well. The systems overall performance is found to be very good and efficient compared to pivot-based system without the sense identification phase.

The paper is organised in such a way that Section 2 gives an overview about the recent works that has been carried out on Indian languages. Section 3 describes about the proposed pivot-based machine translation system. Section 4 discusses about the result outcome of this proposed system and lastly, Section 5 concludes the work and discusses about the future scope of this system.

2 Background work

‘ANUSAARAKAA’ for machine translation from one Indian language such as Telugu, Kannada, Bengali, Punjabi and Marathi to Hindi language or vice versa was developed by IIT Kanpur and IIIT Hyderabad in 1995. This system works on principles of Paninian grammar (PG). The system provides both the robustness and no loss of information. Currently, ANUSAARAKA is working for Telugu, Kannada, Marathi, Bengali and Punjabi to Hindi language translation and in near future reverse translation will also be feasible. The output can be post-edited if there is any grammatical error occurs during machine translation (Bharati et al., 1997). The efficiency of the system was not available.

A machine translating system named ‘MANTRA’ which translates the text from English to Hindi language with a precise domain in office order, administrative work texts, etc. in 1999. The basis of this system was the tree adjoining grammar (TAG) formalism from the University of Pennsylvania. It uses lexicalised tree adjoining grammar (LTAG) for representing the English and the Hindi language. It uses the TAG for parsing as well as generation purposes. Now this system is also used in the finance, agriculture, healthcare, information technology, education and the general purpose activities of the government domains (Darbari, 1999). Currently, the work for the language pairs English-Bengali, English-Telugu, English-Gujarati, Hindi-English, Hindi-Marathi, Hindi-Bengali is also going on. In 2008, the application area of MANTRA was extended to the education and the banking sectors also.

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

Imam et al. (2011) have discussed about the impact of corpus size in English-Bangla statistical machine translation. The author has identified that increase in corpus size to improve translation quality will saturate the quality of translation at particular instant and then, there comes the need for improving quality of the corpus. Author has developed their own corpus from various sources and has evaluated the machine translation system based on BLEU score.

In 2014, the research on statistical machine translation between English and Mauritian Creole language pairs has been developed specifically for tourism and business purpose (Sukhoo et al., 2014). The author has used MOSES tool to develop the system and found that system performance was not up to the mark as the parallel corpus was too small. Author has also used the bilingual dictionary to improve its performance and based on the system’s evaluation they has found that BLEU score to be approximately 6.0. The BLEU score seems to increase with respect to the increase in corpus size.

Wang et al. (2015) have proposed a novel neural network bilingual language model for statistical machine translation system. The continuous space language model makes use of monolingual corpus and the author has proposed a method to modify it to bilingual continuous space language model. They have used the different language models over Chinese-English machine translation system and have used 1 million parallel training data. Author has also suggested the way to reduce the computational and space complexity in processing the identified phrases by considering only top phrases which are identified using ranking method. This proposed system is found to outperform the existing language models converting/growing methods in the statistical machine translation systems.

In 2015, Xiong et al. have proposed a document level topic-based coherence model for statistical machine translation. This system extracts the coherence chain from the source text and this extracted chain is further mapped on to the target coherence chain using maximum entropy classifier. Author has developed two topic-based coherence models over these generated target coherence chain – word level coherence model and the phrase level coherence model. The author has found that the developed system outperforms compared to the baseline system. It is also identified that the phrase level coherence model is comparatively better over word level coherence model.

Saini and Sahula (2015) have surveyed on various machine translation system for Indian languages and have suggested that transfer-based approach is more flexible for multilingual languages. The author has identified the short comings of these existing systems in terms of the immature rule set, dictionary and methodology. It is also identified that the systems that are developed are mostly for Hindi language and very few for south Indian languages. Thus author suggests that there is scope for much development in machine translation system on these languages to reduce the language barriers.

Shahnawaz and Mishra (2015) have presented English to Urdu machine translation system which makes use of case-based reasoning (CBR) to identify the translation rules and these translation rules are further used in the artificial neural network. The author has found that the system's performance is good for those sentences for which the case available in case base.

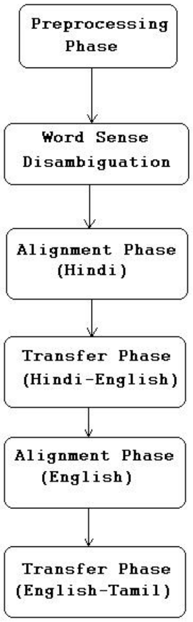
Bakhouche et al. (2015) have proposed an ant colony optimisation algorithm for word sense disambiguation of Arabic words using the lexical information of corresponding English words in Princeton WordNet. The Arabic WordNet has been mapped with the Princeton WordNet and this mapping is used to identify the lexical information of the word under consideration. The performance of this system is found to be approximately 80%.

Soltani and Faili (2012) have developed a system which generates the semantic dependency graph of different sense of word and ranks each node in the graph. This semantic dependency graph is used to identify the target word in English to Persian machine translation system. This system also makes use of statistical information to identify the target word.

3 Proposed system – pivot language-based machine translation

Proposed system is a word level statistical translation system and it has various phases which are as shown in Figure 1. This system has a preprocessing phase in which the Hindi-English-Tamil pairs are mapped and are aligned based on their part-of-speech. With the help of the aligned word pairs the Hindi texts are translated to English in the transfer phase 1 which is followed by transfer phase 2 which translates the pivot language (English) to the target language (Tamil).

Figure 1 Overall block diagram of the system


Preprocessing Phase
Word Sense Disambiguation
Alignment Phase (Hindi)
Transfer Phase (Hindi-English)
Alignment Phase (English)
Transfer Phase (English-Tamil)

3.1 Word alignment phase

This is the major phase of a machine translation system which aligns the words with its corresponding word's position in the other language due to difference in grammar of languages under consideration. There are various algorithms to perform the word alignment process. In this proposed system, the modified IBM model is being used to map between Hindi-English pairs and English-Tamil pairs. The basic IBM model does not consider the part-of-speech of the words during alignment process. This modified IBM model performs the alignment based on the part-of-speech of the word pairs. As per IBM models, the probability of words position in the target language is dependent on the position of words in source language, length of source and target texts. In this modified IBM model, the part-of-speech of the source text is also considered as a parameter for the alignment probability, which is mathematically represented as mentioned in below equation.

$$P\left(\frac{j}{i, l, m, pos}\right) = \frac{P(j, i, l, m, pos)}{P(i, l, m, pos)}$$

where

j position of target word

i position of source word

l length of source text

m length of target text

pos part-of-speech of the source word.

Using probability theory the above equation is re-written as,

$$P\left(\frac{j}{i, l, m, pos}\right) = \frac{P\left(\frac{i}{j, l, m, pos}\right) * P(j, l, m, pos)}{P(i, l, m, pos)}$$

$$= \frac{P\left(\frac{i}{j, l, m, pos}\right) * P\left(\frac{l}{j, m, pos}\right) * P(j, m, pos)}{P(i, l, m, pos)}$$

$$= \frac{P\left(\frac{i}{j, l, m, pos}\right) * P\left(\frac{l}{j, m, pos}\right) * P\left(\frac{m}{j, pos}\right) * P(j, pos)}{P(i, l, m, pos)}$$

$$= \frac{P\left(\frac{i}{j, l, m, pos}\right) * P\left(\frac{l}{j, m, pos}\right) * P\left(\frac{m}{j, pos}\right) * P\left(\frac{pos}{j}\right) * P(j)}{P(i, l, m, pos)}$$

Since Indian languages such as Hindi, Tamil are free word order languages and also the length of source as well as target text are independent events with respect to position of target word, part-of-speech of source word, the above equation is reduced to

$$P\left(\frac{j}{i, l, m, pos}\right) = \frac{P\left(\frac{i}{j, l, m, pos}\right) * P(l) * P(m) * P(pos) * P(j)}{P(i, l, m, pos)}$$

$$P(i, l, m, pos) = P\left(\frac{pos}{i, l, m}\right) * P(i, l, m)$$

$$= P\left(\frac{pos}{i, l, m}\right) * P\left(\frac{m}{i, l}\right) * P(i, l)$$

$$= P\left(\frac{pos}{i, l, m}\right) * P\left(\frac{m}{i, l}\right) * P\left(\frac{l}{i}\right) * P(i)$$

By neglecting independent events, this equation is further reduced to

$$P(i, l, m, pos) = P\left(\frac{pos}{i, l}\right) * P\left(\frac{m}{l}\right) * P(l) * P(i)$$

$$P\left(\frac{j}{i, l, m, pos}\right) = \frac{P\left(\frac{i}{j, l, m, pos}\right) * P(l) * P(m) * P(pos) * P(j)}{P\left(\frac{pos}{i, l}\right) * P\left(\frac{m}{l}\right) * P(l) * P(i)}$$

$$P\left(\frac{j}{i, l, m, pos}\right) = \frac{P\left(\frac{i}{l}\right) * P(m) * P(pos) * P(j)}{P\left(\frac{pos}{i, l}\right) * P\left(\frac{m}{l}\right) * P(i)}$$

The probabilities on the right hand side of the above equation can be calculated based on the corpus that is being used for training purpose. For generating the table that will be required for training, MGIZA++ tool is being used which is a part of the MOSES tool (Koehn et al., 2007). MGIZA++ is basically used for generating alignment table, translation table and so on using the IBM models and HMM. This tool accepts the untagged parallel corpus. But since there is need for modification discussed above, the training document for this tool is made as tagged parallel corpus.

3.2 Semantic analysis

In order to identify the words sense based on the context in which it is being used, there is need for word sense disambiguation. There are various methods that can be used for word sense disambiguation such as point mutual information, latent semantic analysis (LSA), etc. LSA is used in this proposed system since LSA helps to identify the semantic relationship between different words. LSA is a matrix decomposition method which decomposes the term frequency matrix in to left singular matrix, right singular matrix and singular diagonal matrix (Kumar et al., 2015). Each of these three matrices interprets about certain features of the language such as the left singular matrix represents the words and their relationship with the documents under consideration. Similarly, the relationship between documents and the words are represented in the right singular matrix and the singular diagonal matrix is the representation of the words in documents which are more semantically equivalent. The dot product of these left singular matrix and the singular diagonal matrix basically gives the words semantic representation which can be used for identifying the sense in which it is being used in that sentence. LSA is mathematically represented as below,

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

where f_{tn}^{m} indicates the frequency of nth term in mth sentence.

In this proposed system, the term frequency matrix is generated from the input sentence and sentences that are retrieved from Hindi WordNet which has same words as in input but has different senses. This generated term frequency matrix is subjected to matrix decomposition which further generates three different matrix – left singular matrix, right singular matrix and singular diagonal matrix. The dot product of these generated left singular matrix and singular diagonal matrix will give the most related sense of the word in that particular context.

3.3 Transfer phase

This proposed system has two transfer phases since there is a use of pivot language during the translation process which is represented as

$$P\left(\frac{w_t}{w_s,\ pos}\right) = P\left(\frac{w_p}{w_s,\ pos}\right) * P\left(\frac{w_t}{w_p,\ pos}\right)$$

It is clear from the above equation that the translation from source to pivot has to be performed first and the other translation will be on the basis of the probable pivot word identified. Thus, initial transfer phase is for the translation from source to pivot whereas the second one is for translation from pivot to target language. The detailed descriptions about these phases are mentioned as follows.

3.3.1 Transfer phase 1 (Hindi to English)

Before discussing about the methodology, let us discuss about the relationship between these languages which are under consideration. Basically, the English language is a fixed word order language whereas the Hindi language is a free word order language. During the analysis it is found that there a one-to-one transfers, many-to-one transfer, one-to-many transfer required for these languages. The one-to-one transfer is direct word by word translation which does not have much complicacy. But during the translation from one-to-many and many-to-one there is a need for certain methods to improve its accuracy. In this proposed system, one-to-many translation has been taken care of by using the part-of-speech and the semantics of the input word. For the many-to-one translation, there is need for statistical information about the words that can be grouped during the translation process. For this purpose, the system makes use of *n*-gram statistical analysis where *n* consecutive words are considered during the transfer phase. The statistical transfer from source to pivot is represented mathematical as follows,

$$P\left(\frac{w_p}{w_s,\ pos}\right) = P\left(\frac{w_s,\ pos}{w_p}\right) * P\left(w_p\right)$$

Since the part-of-speech of source word and the source word are independent event on the condition of the pivot word for it, the above equation is reduced to,

$$P\left(\frac{w_s,\ pos}{w_p}\right) = P\left(\frac{w_s}{w_p}\right) * P\left(\frac{pos}{w_p}\right)$$

$$P\left(w_p\right) = P\left(\frac{w_p}{w_{p-1}}\right)$$

3.3.2 Transfer phase 2 (English to Tamil)

This phase also works in similar manner as it was done in previous source-pivot transfer phase. The target language (i.e., Tamil) is also free word order language which is on contrary to the English language and both of the languages follow a different grammar.

The word alignment phase is once again needed in this phase of translation to ensure the grammatical correctness of the target language. Thus there will be need for alignment table as well as a translation table, which are generated using the MGIZA++ tool with a slight modification on to the table. In general, the alignment table has the position of target word (*j*), position of pivot word (*i*), length of pivot text (*l*), length of target text (*m*) and probability (*p*(*j*/*i*, *l*, *m*)). The alignment table in this proposed system has been modified by including part-of-speech of the pivot word in addition to all the other parameters. The translation table has the pivot word (w_p), part-of-speech of the pivot word (*pos*), its equivalent target word (w_t) and their respective probability. This transfer phase can be represented mathematically as,

$$P\left(\frac{w_t}{w_p,\ pos}\right) = P\left(\frac{w_p}{w_t}\right) * P\left(\frac{pos}{w_t}\right) * P\left(\frac{w_t}{w_{t-1}}\right)$$

4 Results and evaluation

This proposed system has been implemented and the data used for this system (25,000 sentences in each three languages, i.e., Hindi, Tamil and English) has been provided by Technology Development for Indian Languages (TDIL) programme initiated by the Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology (MC&IT), Govt. of India. This whole data has been divided for training and testing purpose. The system has been tested with 10% of data out of the whole corpus and the rest of the data where used for training the system. The data used for evaluating the system has been kept varying to analyse the performance of the system. The system is evaluated and compared with a system which does not have sense identification phase in it. The comparison is as shown in Table 1 and it is evident from this table that the performance of the proposed system is considerably good compared to the other one. The corpus size mentioned in Table 1 is the size of the training data and it is 90% of the data that was being used during analysis. It is found from the analysis that the precision and recall of word sense-based pivot language machine translation has a considerable improvement compared to the other. As there is increase in precision and recall with respect to corpus size, the system has been tested for larger data as well but it is found to decrease the performance due to degradation in the quality of data with the increase in corpus size. In Table 2, the comparison between pivot-based Hindi-Tamil machine translation and direct Hindi-Tamil machine translation has been shown. It is being noted from Table 2, that the performance of word sense-based Hindi-Tamil machine translation using pivot language has improved but there is decrease in the percentage of recall. This decrease in recall is due to the distortion that occurs because of the inclusion of the pivot language in between. This distortion keeps increasing with respect to the increase in corpus size as the noise in the corpus also gets increased.

Table 1 Comparison of pivot-based Hindi-Tamil machine translation with/without word sense

S. no.	*Corpus size (in number of words)*	*Pivot-based Hindi-Tamil machine translation*		*Word sense-based Hindi-English-Tamil machine translation*	
		Precision (in %)	*Recall (in %)*	*Precision (in %)*	*Recall (in %)*
1	10,000	53	52	68	58
2	20,000	64	67	70	69
3	30,000	76	74.5	81	76

Table 2 Comparison of pivot-based machine translation with direct Hindi-Tamil machine translation

S. no.	*Corpus size (in number of words)*	*Word sense-based Hindi-English-Tamil machine translation*		*Word Sense-based Hindi-Tamil machine translation*	
		Precision (in %)	*Recall (in %)*	*Precision (in %)*	*Recall (in %)*
1	10,000	68	58	65	63
2	20,000	70	69	76	72.5
3	30,000	81	76	87	86.5

Figures 2 and 3 show the comparison of various systems in terms of precision and recall respectively. It is very clear from Figure 2 that the precision of proposed word sense-based Hindi-English-Tamil MT system improves with respect corpus size, but its precision is lesser when compared with word sense-based Hindi-Tamil MT system which is due to the increase in distortion by the inclusion of pivot language during translation. From Figure 3, it is visible that the recall is relatively good compared to the pivot-based Hindi-Tamil MT but degrades when compared to the word sense-based Hindi-Tamil MT due to the distortion.

Figure 2 Comparison of MT systems in terms of precision (see online version for colours)

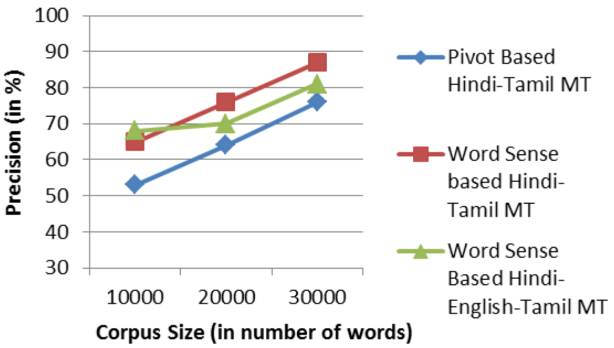
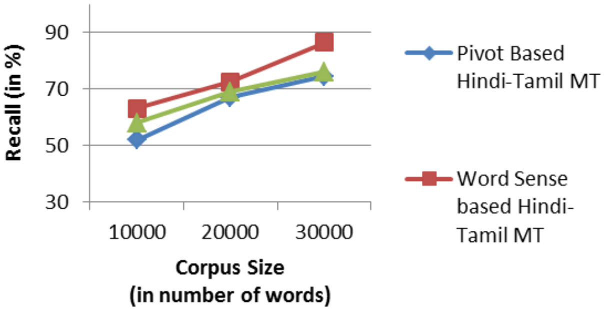


Figure 3 Comparison of MT systems in terms of recall (see online version for colours)



Since there are certain shortcomings in naïve Bayes statistical machine translation method, this proposed system has been incorporated with certain modifications to cope up with the shortcomings which have improved the performance of the system considerably. Table 3 lists out the shortcomings identified in naïve Bayes statistical machine translation and the proposed modification in the system.

Table 3 Comparison of proposed machine translation with naïve Bayes statistical machine translation

S. no.	*Feature*	*Naïve Bayes statistical machine translation*	*Advantage/ disadvantage*	*Proposed machine translation*	*Advantage/ disadvantage*
1	Language model $P(s)$	Used for source language	It considers preceding word during translation mechanism which improves the translation	Used for source language as well as pivot language	It considers preceding word during translation and pivot language is used as there is low resource availability
2	Translation model $P(s\|t)$	Used for target language	It maps the source word with the target language and predicts the probable translation which also helps during translation	Used for target as well as pivot language	It has the same advantage as in naïve Bayes but used pivot since the languages are low resource language
3	Part-of-speech (POS)	Not used	It is a disadvantage	Used in the translation model which is modified accordingly	POS provides more accurate mapping between source and target words. Thus it improves the translation accuracy

Table 3 Comparison of proposed machine translation with naïve Bayes statistical machine translation (continued)

S. no.	*Feature*	*Naïve Bayes statistical machine translation*	*Advantage/ disadvantage*	*Proposed machine translation*	*Advantage/ disadvantage*
4	Alignment model	Used for word alignment	Uses traditional IBM models and does not consider POS during alignment	Used for word alignment but has been modified to consider part-of-speech during alignment process	Uses modified IBM model and its alignment improves since the word position depends on the POS of the word as well
5	Word Sense	Since it works only on probabilistic manner, it does not include sense identification.	Since contextual information is not considered translation is poor in certain cases	During translation phase, it considers the words sense in the context it is being used	Improves the translation based on contextual information. Words translation differs based on the sense it is being used

BLEU Score is bilingual evaluation understudy which is used to evaluate the accuracy of evaluation when compared with its reference sentences (Papineni et al., 2002). The proposed system is found to have a BLEU score of 0.7637. The BLEU score of system without word sense identification is found to be 0.7394 and thus there is an improvement in translation by 0.03 when word sense is being considered. Table 4 compares the proposed system with existing statistical machine translation systems (in other languages) and the proposed system is found to have good BLEU score.

Table 4 Comparison of BLEU Score with various statistical machine translations

S. no.	*Methodology*	*Source language*	*Target language*	*BLEU score*
1	Statistical machine translation (Mantoro et al., 2013)	English	Bahasa Indonesia	0.2287
2	Lemma translation (Sulaeman and Purwarianti, 2015)	Japanese	Indonesian	0.1282
3	Lemma translation (Sulaeman and Purwarianti, 2015)	Indonesian	Japanese	0.1723
4	Proposed pivot-based translation	Hindi	Tamil	0.7394
5	Proposed pivot-based word sense translation	Hindi	Tamil	0.7637

5 Conclusions and future work

This proposed system has good accuracy when compared with the system without sense disambiguation phase. But when compared with system which does not use a pivot language, the system has degradation in its performance this is due to the increase in noise due to the use of English language as pivot. The noise also gets doubled with the increase in corpus size. Thus the system performance keeps decreasing if there is increase in corpus size. Ideally, it has been identified that the corpus size can be kept at 30,000.

There can be a good improvement in the system's performance if the quality of data is improved with the increase in corpus size. Use of quality data can help in reducing the distortion that is occurring due to the usage of pivot language in between. It can also improve the performance by using a pivot language which is syntactically/semantically related to both the source and target language. For example, in case of Hindi and Tamil language, Sanskrit language can be used a pivot. But only bottleneck is the resource availability in Sanskrit. This system can also be extended further by using some other approach of word sense disambiguation.

References

Bakhouche, A., Yamina, T., Schwab, D. and Tchechmedjiev, A. (2015) 'Ant colony algorithm for Arabic word sense disambiguation through English lexical information', *Int. J. Metadata Semant. Ontologies*, December, Vol. 10, No. 3, pp.202–211.

Bharati, A., Vineet, C., Kulkarni, P.A. and Sangal, R. (1997) 'Anusaaraka: machine translation in stages', *A Quarterly in Artificial Intelligence*, July, Vol. 10, No. 3, pp.22–25, NCST, Mumbai.

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L. (1993) 'The mathematics of statistical machine translation: parameter estimation', *Computational Linguistics*, Vol. 19, No. 2, pp.263–311.

Darbari, H. (1999) 'Computer-assisted translation system – an Indian perspective', *Machine Translation Summit VII*, 13–17 September, Kent Ridge Digital Labs, Singapore, pp.80–85.

Imam, A.H., Arman, M.R.M., Chowdhury, S.H. and Mahmood, K. (2011) 'Impact of corpus size and quality on English-Bangla statistical machine translation system', *2011 14th International Conference on Computer and Information Technology (ICCIT)*, 22–24 December, pp.566–571.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007) 'Moses: open source toolkit for statistical machine translation', *Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, June, Prague, Czech Republic.

Kumar, K.V., Yadav, D. and Kumar, A. (2015) 'Graph based technique for Hindi text summarization', *Second International Conference on Information Systems Design and Intelligent Applications – 2015, Advances in Intelligent and Soft Computing (AISC)*, 8–9 January, Vol. 339, pp.301–310, Springer, ISBN: 978-81-322-2249-1, ISSN:2194-5357.

Mantoro, T., Asian, J., Octavian, R. and Ayu, M.A. (2013) 'Optimal translation of English to Bahasa Indonesia using statistical machine translation system', *2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, 26–27 March, pp.1–4.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) 'BLEU: a method for automatic evaluation of machine translation', *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pp.311–318, Stroudsburg, PA, USA.

Saini, S. and Sahula, V. (2015) ‘A survey of machine translation techniques and systems for Indian languages’, *2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT)*, 13–14 February, pp.676–681.

Shahnawaz and Mishra, R.B. (2015) ‘An English to Urdu translation model based on CBR, ANN and translation rules’, *Int. J. Adv. Intell. Paradigms*, July, Vol. 7, No. 1, pp.1–23.

Comment [t1]: Author: Please provide the author’s name with first name initials.

Sinha, R.M.K. (2004) ‘An engineering perspective of machine translation: AnglaBharti-II and AnuBharti-II architectures’, *Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS-2004)*, 17–19 November, pp.134–138, Tata McGraw Hill, New Delhi.

Soltani, M. and Faili, H. (2012) ‘Target word selection in English to Persian translation using unsupervised approach’, *Int. J. Artif. Intell. Soft Comput.*, September, Vol. 3, No. 2, pp.125–142.

Sukhoo, A., Bhattacharyya, P. and Soobron, M. (2014) ‘Translation between English and Mauritian Creole: a statistical machine translation approach’, *IST-Africa Conference Proceedings 2014*, 7–9 May, pp.1–10.

Sulaeman, M.A. and Purwarianti, A. (2015) ‘Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process’, *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, 10–11 August, pp.54–58.

Wang, R., Zhao, H., Lu, B-L., Utiyama, M. and Sumita, E. (2015) ‘Bilingual continuous-space language model growing for statistical machine translation’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, July, Vol. 23, No. 7, pp.1209–1220.

Xiong, D., Zhang, M. and Wang, X. (2015) ‘Topic-based coherence modeling for statistical machine translation’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, March, Vol. 23, No. 3, pp.483–493.

Bibliography

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.*

Bharati, A., Chaitanya, V. and Sangal, R. (1991) ‘Local word grouping and its relevance to Indian languages’, in Bhatkar, V.P. and Regeedn, K.M. (Eds.): *Frontiers in Knowledge Based Computing (KBCS90)*, Narosa Publishing House, New Delhi.

Goyal, V. (2010) *Development of a Hindi to Punjabi Machine Translation System*, October [online] http://www.languageinindia.com.

Comment [t2]: Author: Please provide the access details (date when the site was accessed/visited).

Goyal, V. and Lehal, G.S. (2008) ‘Hindi morphological analyzer and generator’, *Proceedings of International Conference on Emerging Trends in Engineering and Technology (ICETET ‘08)*, pp.1156–1159.

Goyal, V. and Lehal, G.S. (2009) ‘A machine transliteration system for machine translation system: an application on Hindi-Punjabi language pair’, *Atti Della Fondazion Giorgio Ronchi (Italy)*, Vol. 64, No. 1, pp.27–35.

Goyal, V. and Lehal, G.S. (2010) ‘Web based Hindi to Punjabi machine translation system’, *Journal of Emerging Technologies in Web Intelligence*, May, Vol. 2.

Comment [t3]: Author: Please provide the issue number and page numbers.

Huang, C-C., Chen, M-H., Yang, P-C. and Chang, J.S. (2013) ‘A computer-assisted translation and writing system’, *ACM Transaction on Asian Language Information Processing*, October, Vol. 12, No. 4, Article 15, 20pp.

Comment [t4]: Author: Please provide the page numbers.

Korra, R., Sujatha, P., Chetana, S. and Kumar, M.N. (2011) ‘Performance evaluation of multilingual information retrieval (MLIR) system over information retrieval (IR) system’, *IEEE-International Conference on Recent Trends in Information Technology (ICRTIT).*

Kumar, P. and Goyal, V. (2010) 'Development of Hindi-Punjabi parallel corpus using existing Hindi-Punjabi machine translation system and using sentence alignment', *International Journal of Computer Applications*, Vol. 5, No. 9.

Pandian, S.L. and Geetha, T.V. (2008) 'Morpheme based language model for Tamil part-of-speech tagging', *Research Journal on Computer Science and Computer Engineering with Applications (POLIBITS08)*, Vol. 38, pp.19–25.

Paul, M., Finch, A. and Sumita, E. (2013) 'How to choose the best pivot language for automatic translation of low-resource languages', *ACM Transaction on Asian Language Information Processing*, October, Vol. 12, No. 4, Article 14, 17pp.

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

Saraswati, J., Shukla, R., Goyal, R.P. and Bhattacharyya, P. (2010) 'Hindi to English Wordnet linkage: challenges and solutions', *8th International Conference on Natural Language Processing*.

Saravanan, K., Parthasarathi, R. and Geetha, T.V. (2003) 'Syntactic parser for Tamil', *Proceedings of the Tamil Internet Conference*, Chennai, Tamilnadu, India, pp.28–37.

Sinha, R.M.K. and Thakur, A. (2005) 'Machine translation of bi-lingual Hindi-English (Hinglish) text', *Proceedings of the Tenth Machine Translation Summit, MT Summit X*, 13–15 September, Phuket, Thailand, pp.149–156.

Yadav, R.K. and Gupta, D. (2010) 'Annotation guidelines for Hindi-English word alignment', *International Conference on Asian Language Processing*.

Comment [t5]: Author: Please provide the page numbers.

Comment [t6]: Author: Please provide the issue number.