Rule based Approach for Prepositional Phrase Attachment in English-Tamil Translation

S.Suganthi
Dept of CSE-PG
National Engineering College
K.R.Nagar,Kovilpatti
Tamil Nadu,India.

P.Bama Ruckmani
Asst.Professor/CSE-PG
National Engineering College
K.R.Nagar,Kovilpatti
Tamil Nadu,India.

K.G.Srinivasagan, PhD.
Prof. & Head/CSE-PG
National Engineering College
K.R.Nagar,Kovilpatti
Tamil Nadu,India.

M.Saravanan
Dept of CSE/PG
National Engineering College
K.R.Nagar,Kovilpatti
Tamil Nadu,India.

ABSTRACT
This paper mainly under the field of Natural Language Processing (NLP). Machine Translation is a major application area under NLP. The main aim of this work is to improving the translation quality especially in English-Tamil. There are so many researchers are already developed their work in this field. But the human expectation is not yet achieved. So the translation research is still exist. While translating English-Tamil translation prepositional phrase attachment and orthographic errors are the major issues. Different kinds of prepositions are used quite normally in English, in context of Tamil translation focusing towards English prepositions alone. Always English prepositions are treated as postposition in Tamil. Place the postposition is based on 'time, place, direction, context'. In some context different preposition may promote unique meaning, in such scenarios the Word Sense Ambiguation problem may arise. To resolve the Word Sense Ambiguation and word reordering, an algorithm called "Prepositional Phrase Attachment" is proposed. This system handles the frequently used prepositions such as "of, in, to, on, by, from". The correct meaning of a prepositional word is achieved through this work.

Keywords
Machine Translation, Prepositional Phrase Attachment, Orthographical Rules, POS tagging, Words Reordering.

1.INTRODUCTION
The language has been playing a major role in all the sectors. For the purpose of communicating with the world wide people, and accessing scientific resources in the major field the language knowledge is needed. Literacy in the mother tongue is no longer enough to follow the information supplied by the other languages. so it is necessary to bridge the gap with the help of modern technologies as early as possible. In this context the machine translation is essential. The main issue of the translation is structural order of the sentence. Language sentences have many parts of speech like subject, verb and object. Structural order of language is very from language to language. The words reordering might be difficult and important task because the order of words may affect the original meaning of a sentence. Basically English and Tamil languages are different structural order. The following fig.1 illustrates the structural order of English and Tamil language.

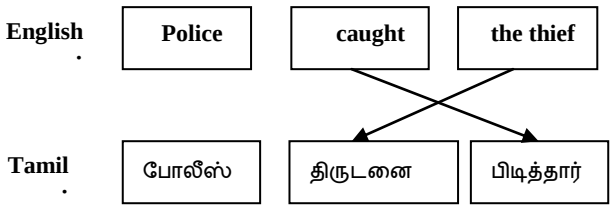

Figure 1. The structural order of English and Tamil

In the existing system (Google Translate) already deal with solution for the problem of machine translation of almost 40 languages. Especially in English-Tamil translation accuracy is not appreciable in many cases, particularly while dealing with the preposition and spelling errors. In order to obtain the correct English to Tamil translation the correct preposition, correct spelling and word structure is very essential. The rest of the paper is organized as follows, the related works are discussed in section 2, framework and proposed algorithm presented in section 3, experimental results and discussion are reported in section 4 and followed by concluding remarks.

2. RELATED WORK
Earlier many researchers have concentrated in language translation using prepositions, idioms, phrasal verbs, and converting a complex sentence into simple sentences. They obtained and reported the different levels of accuracy as result. The results may not fulfill the scenario of users need.

Micheal Collins et al [1] proposed a Backed off method for handling the prepositional phrase attachments, which is appreciably better than other methods which have been tested on the wall street journal corpus. Their algorithm has the additional advantages of being conceptually simple and computationally inexpensive to implement. They reported as accuracy of 84.5% is close to the human performance figure of 88% using the 4 head words alone. Sanda M.Harabagiu [2] presented a method for word sense disambiguation and coherence understanding of prepositional relations. This method is to classify prepositional attachments according to semantic equivalence of phrase heads and then apply

inferential heuristics for understanding the validity of prepositional structures. This paper proposes a method of extracting and validating semantic relations for prepositional attachment the method.

Sudip Kumar Naskar et al [4] presented the approach of handling of English prepositions in Bengali has been studied with reference to a machine translation system from English to Bengali. In machine translation, sense disambiguation of preposition is necessary when the target language has different representations for the same preposition. In Bengali, the choice of the appropriate inflection depends on the spelling of the reference object.The choice of the postpositional word depends on the semantic information about the reference object obtained from the WordNet. V. Dhanalakshmi et al [5] presented the grammar teaching tools for analyzing and learning character, word and sentence of Tamil language. Tools like Character Analyzer for analyzing character, morphological Analyzer and generator and verb conjugator for the word level analysis and parts of speech tagger, Chunker and dependency parser for the sentence level analysis were developed using machine learning based technology.

Poornima C et al [6] proposed a rule based technique for simplifying the complex sentences into simple sentences based on connectives like pronouns, coordinating and subordinating conjunction without changing the meaning of the sentence. This method is useful as a preprocessing tool for machine translation. It has been proved that the splitting technique can lead to remarkable improvements in machine translation system. Dr.S.Saraswathi et al [7] developed a bilingual translation system for English and Tamil using hybrid approach. They use Rule based machine translation (RBMT) and Knowledge based machine translation (KBMT) techniques. New rules have been added to the proposed system in order to make the system more efficient.

Matt Post et al [8] described the collection of six parallel corpora containing four-way redundant translations of the source-language text. The Indian languages of these corpora are low-resource and understudied, and exhibit markedly different linguistic properties compared to English. They performed baseline experiments and suggested a number of approaches that could improve the quality of models constructed from the datasets. S. Lakshmana Pandian et al [9] presents an effective methodology for English to Tamil translation. They implemented in a Rule based approach which involves segmentation and tagging, Rule based Reordering, Morphological Analyzing and dictionary based translation to the target language. Then the errors in the translated sentences are corrected by applying Statistical technique. Since a word in English has multiple meaning in Tamil an effective word dictionary file is needed in order to achieve better results in translation.

P G Thiruumeni et al [10] provides a technique for used to handle the idioms and phrasal verbs during the translation process and it increases the accuracy of the translation. The BLEU and NIST scores calculated before and after handling the phrasal verbs and idioms during the translation process show a significant increase in the accuracy of the translation. The proposed technique for used in English to Tamil machine translation system, can be incorporated with machine translation system for English to any language. This approach can be used in both rule based and factored statistical machine translation with some modifications.

3. PREPOSITIONAL PHRASE ATTACHMENT

Syntactically prepositions can be arranged into three classes- simple prepositions, compound prepositions and phrase prepositions. Different kinds of prepositions are used quite normally in English, in context of Tamil translation focusing towards English prepositions alone. Always English prepositions are treated as postposition in Tamil. Place the postposition is based on 'time', 'place or direction' and 'context'. In some context different prepositions may promote unique meaning, in such scenarios the Word Sense ambiguation problem may arise. To resolve the Word Sense Ambiguation problem and word reordering, an algorithm called "***prepositional attachment algorithm"*** is proposed.

The Rule-Based approach is used in the proposed algorithm to solve the above mentioned problem. The correct meaning of a preposition is chosen by using a rule-based approach and is placed incorrect position using the semantic structure of a target language. Our system handles the frequently used prepositions such as ***'of ', 'in', 'to', 'on', 'by', 'from'.*** The semantic rules are fully based on parts of speech in a given sentence. The PENN tree tag is used for the purpose of assigning the POS tag. The PENN tree tag contains approximately 36 tags with examples. In order to overcome word sense disambiguation the proposed algorithm extracts the triplet words **(*Preceding word of preposition, Preposition, Succeeding word of preposition).***

The proposed algorithm illustrates the following steps:

1. Segmenting the sentences from the larger paragraph based on delimiters such as ".", "?".
2. Assign POS tag for every word in a given sentence using PENN tree grammar set.
3. Check whether the prepositions is presence or not in a current sentence.
4. If the preposition is found then to extract the triplet tag.
5. Decision based on the middle word of the triplet.
6. The words are reordered based on the semantic structure of the target language.

The text can be in any form either individual sentences or paragraph format. The simple sentences are converted from the paragraph using the delimiters such as ***(".", "?").*** The simple sentence obtained in each word is assigned as the parts of speech using the PENN tree grammar set. Based on the POS tag the rules are generated.

POS (Parts of Speech) is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context. The POS is useful for assigning the grammatical categories or word category disambiguation for each and every English words in English to Tamil translation. In this paper, taken as a PENN Tree Tag set. This tag set nearly 36 tags. The following table. I illustrates some of the example of PENN Tree POS (Parts Of Speech) tag set.

Rules of the prepositional phrase "of "

1. If the order of triplets presents in a given sentence is <NN><IN><DT> or <NN><IN><NN> then the meaning

of a prepositional phrase should be **"udaya/in**"[**உடைய/இன்**].

2. If the order of triplets presents in a given sentence as <NN><IN><JJ> then the meaning of a prepositional phrase is given as **"**kkhana"[**க்கான]**

3. If the order of triplets presents in a given sentence as <RB><IN><NNP> then the meaning of a prepositional phrase is given as **"**ill" [**இல்**].

4. If the order of triplets presents in a given sentence as <VBN><IN><IN> then the meaning of a prepositional phrase is given as **"aal**"[**ஆல்**].

Rules of the prepositional phrase "from "

If the order of triplets presents in a given sentence is <POSP1><IN><POSP2> then the meaning of a prepositional phrase is **"irunthu".**

Rules of the prepositional phrase "by"

If the order of triplets presents in a given sentence is <POSP1><IN><POSP2> then the meaning of a prepositional phrase is **"aal".**

Rules of the prepositional phrase "on"

If the order of triplets presents in a given sentence is <POSP1><IN><POSP2> then the meaning of a prepositional phrase is **"mele/il".**

Rules of the prepositional phrase "in"

If the order of triplets presents in a given sentence is <POSP1><IN><POSP2> then the meaning of a prepositional phrase is **"il".**

Rules of the prepositional phrase "to"

If the order of triplets presents in a given sentence is <POSP1><IN><POSP2> then the meaning of a prepositional phrase is **"kku".**

The proposed system is designed to handle the prepositions such as ***'of ', 'in', 'to', 'on', 'by', 'from'.*** The following table 1 shows the Tamil meaning of these prepositions.

S.No	Preposition	Meaning in Tamil
1	of	1. udaiya [உடைய] 2. in[இன்] 3. kkaana[க்கான] 4. aal[ஆல்] 5. il [இல்]
2	in	il [இல்]
3	to	Kku[க்கு]
4	on	Mele[மேலே]
5	by	Aal[ஆல்]
6	from	Irunthu[இருந்து]

Table 1. Tamil meaning of prepositions *'of ', 'in', 'to', 'on', 'by', 'from'*

The detailed algorithm for prepositional phrase attachment is as follows:

Algorithm for attaching the Prepositional phrase "of "

Let the paragraph be **"P"** and split into **"S_1", "S_2", "S_3"…."S_n"**
Take s_1 into the number of segments **"W_1", "W_2", "W_3"…."W_n"**
For i = 1 to n
$W_1 \ldots W_n$ <- **<Pos tag>**
From (1 to n)
Check the presence of "Preposition" or "Not"
If so Flag=1
Else
Flag=0
End if
If(flag==1)
For i=1 to n
Extract the Triplet term <$POS_{(P1)}$><IN><$POS_{(P2)}$> and store in "**T**"
If ((<IN> == "of ") && (<$POS_{(P1)}$>==<NN>) && (<$POS_{(P2)}$>==<NN>)) ||
((<IN> == "of") && (<$POS_{(P1)}$>==<NN>) && (<$POS_{(P2)}$>==<DT>))
Then
<IN> ← 'udaiya/in'
T <- <POS $_{(P2)}$>||<IN>||<POS $_{(P1)}$>
Main Phrase ← T
Else
If ((<IN> == "of") && (<$POS_{(P1)}$>==<NN>) && (<$POS_{(P2)}$>==<JJ>))
Then
<IN> ← 'kkaana'
T <- <POS $_{(P2)}$>||<IN>||<POS $_{(P1)}$>
Main Phrase ← T
Else
If ((<IN> == "of") && (<$POS_{(P1)}$>==<RB>) && (<$POS_{(P2)}$>==<NNP>))
Then
<IN> ← 'il'

T <- <POS $_{(P2)}$>||<IN>||<POS $_{(P1)}$>
Main Phrase ← T
Else
If ((<IN> == "of") && (<POS$_{(P1)}$>==<VBN>) && (<POS$_{(P2)}$>==<NN>))
Then
<IN> ← 'aal'
T <- <POS $_{(P2)}$>||<IN>||<POS $_{(P1)}$>
Main Phrase ← T
End if
End if
End if
End if
If(flag ==0)
Main Phrase← ($w_1||w_2\ldots||w_n$)
End if;
End **for.**

4. ORTHOGRAPHICAL RULES

An orthography is a standard system for using a particular writing system for any language. The orthographic rules are also standard spelling rules that specify the changes that occur when two morphemes are combined together. An example would be: singular English words ending with -y, when pluralized, end with –ies. The orthography rules are language dependent. Thus these rules have to be framed for each language with the concern of linguistics. It includes rules of spelling, and may also concern other elements of the written language elements such as punctuation and capitalization. If a language uses multiple writing systems, it may have distinct orthographies, as is the case with Kurdish, Uyghur, Serbian, Inuktitut and Turkish. In some cases orthography is regulated by bodies such as language academies, although for many languages (including English) there are no such authorities, and orthography develops through less formal processes. The existing does not concentrate the orthography process during the translation. In order to improve the translation quality, the orthographic rules are essential. The better accuracy can be guaranteed to achieve through this semantic rules. Generally Tamil Nouns ending with

Vowels: ஆ, இ, ஈ, உ, ஊ, ஐ and Consonants: ண்,ம்,ய்,ர்,ல்,ழ்,ள்,ன்

Rule 1:
If the suffix begins in a consonant and the word ends in an consonant ர்,ண்,ம்,ய்,ர்,ல்,ள்,ழ்,ன் then insert '**உ**' in between.

நண்பர்+க்கு => நண்பர்+உ+க்கு
c+c

Rule 2:
If the suffix begins in a vowel and the word ends in an consonant then both the vowels and the consonant are combined together to form new Vowel.

நNNnpaண்பர்+ஆல்=>நண்பரால்
c+ v
லண்டன்+இல்=>லண்டனில்
c +v

Rule 3:
If the suffix begins in a consonant and the word ends in an vowels then add both words without changing any letter.

சென்னை+க்கு=>சென்னைக்கு
v + c

Rule 4:
If the suffix begins in a vowel sound and the word ends in an **இ,ஈ,ஏ,ஐ** insert a **'ய்'** in between.

கவிதை+இல்=>கவிதை+ய்+இல்=>கவிதையில்
c + v

5. EXPERIMENTAL RESULTS

The Proposed frame work and algorithm is experimented with 250 sentences of text from the news papers and articles. All the sentences are used for training. In order to evaluate the system, we applied 250 test sentences, in that 220 sentences are correctly translated. Moreover, Precision and Recall of words are widely used metrics to evaluate the efficiency of Machine translation systems. Precision is nothing but the percentage of generated words that are actually correct. The recall stands for the percentage of words that are generated and that are actually found in the reference translation. F-Measure is the harmonic mean of recall and precision.

$$\text{Precision} = \frac{\text{No. of correctly generated Words}}{\text{Total No. of Words}} = 88\%$$

$$\text{Recall} = \frac{\text{No. of correctly generated Words}}{\text{Total No. of Translated Words}} = 93\%$$

$$\text{F-Measure} = \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall}) / 2} = 90\%$$

The comparative study is made with Google translate. The same set of data are used with Google translate, out of which 130 sentences are correct, the main reason is semantic analyzed of prepositions and reordering error. We obtained the precision, recall and F-measure as 88 %, 93 %, 90 % respectively is as shown in table 2. The Fig. 2 emphasizes very clearly that the proposed system performance is better with respect to all the metric.

Table 2. Experimental analysis of various metrics

System / Metrics	Total. No. of Sentences	Total No. of Words	Total No. of Translated Words	No. of Correct sentences	No. of Correct words	*P(%)	*R(%)	*F(%)
Proposed System	250	1270	1200	220	1120	88	93	90
Google Translate	250	1270	1200	180	960	75	80	77

***P-Precision, *R-Recall, *F-F-Measure**

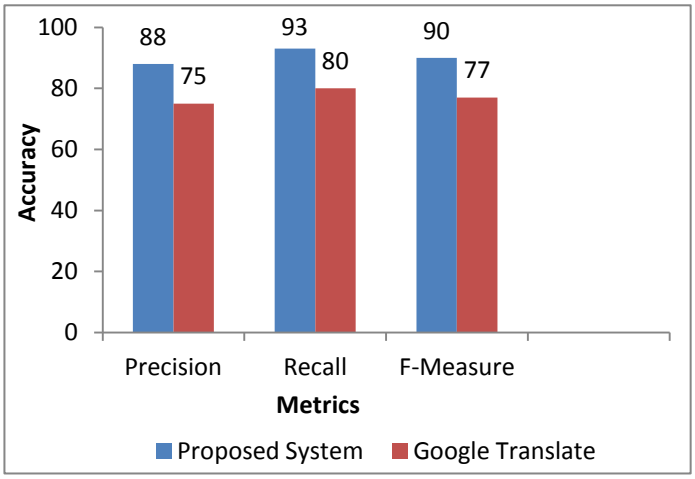

Figure 2. Comparative study of various metrics

6. CONCLUSION

English-Tamil translation using semantic approach for prepositional phrase attachment is implemented in Java environment. In this work, we identified the exact meaning of the preposition with respect to the content and place for English-Tamil translation. The main issue of Word Sense ambiguation is addressed using rule based approach. We experimented. the system and found that the reliability and performance are good. Totally 250 sentences were considered for translation, 220 sentences are of correct translation. Also we calculated Precision, Recall and F-Measure and the corresponding performance is 88%, 93%, 90%. The proposed system was compared with Google Translate and the performances were reported. In future , we planned to concentrate and explore additional idioms and phrases and tense marker approaches and to determine whether the preposition is used in a spatial or temporal sense and also to make it helpful for the task of predicting determiners, prepositions, and other functional words.

6. REFERENCES

[1] Micheal Collins and James Brooks. (1995). "prepositional phrase attachment through a Back-off model". In Proceedings of the Third Workshop on Very Large Corpora, pages 27-38.

[1] Sanda M.Harabagiu. (1996). "An application of wordnet to prepositional attachment". The Association of Computational Linguistics Anthology Network.

[2] I.Dan Melamed, Ryan Green and Joseph P. Turian.(2003) "Precision and Recall of Machine Translation". Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Volume 2, pages 61-63.

[3] Sudip Kumar Naskar and Sivaji Bandyopadhyay. (2006). "Handling of Prepositions in English to Bengali Machine Translation". Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, Trento, Italy, Association for Computational Linguistics, pages 89–94.

[4] Dhanalakshmi V and Rajendran S. (2010). "Natural Language processing Tools for Tamil grammar Learning and Teaching". International journal of Computer Applications (0975-8887), Volume 8, No.14.

[5] Poornima Poornima C, Dhanalakshmi V, Anand Kumar M and Soman K P (2011). "Rule based sentence simplification for English to Tamil Machine Translation System". International journal of Computer Applications (0975-8887),Volume 25, No.8.

[6] Dr.S.Saraswathi , P. Kanivadhana, M. Anusiya and S.Sathiya (2011). "Bilingual Translation System". International Journal on Computer Science and Engineering", Volume 3, No.3.

[7] Matt Post, Chris Callison-Burch and Miles Osborne (2012). "Constructing Parallel corpora for six Indian languages via crowd sourcing". proceedings of the 7th workshop on statistical machine translation, Association for computational linguistics, pages 401-409.

[8] Lakshmana Pandian S and Kumanan Kadhirvelu.(2012). "Machine Translation from English to Tamil using Hybrid Technique". International journal of Computer Applications (0975-8887),Volume 46, No.16.

[9] Thiruumeni P G, Anand Kumar M, Dhanalakshmi V and Soman K P. (2012). "An Approach to Handle Idioms and Phrasal Verbs in English-Tamil Machine Translation System". International Journal of Computer Applications (0975 – 8887), Volume 26, No.10.

[10] Boxing Chen, Roland Kuhn and Samuel Larkin. (2012). "PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pages 930–939.