Pertanika J. Soc. Sci. & Hum. 22 (4): 1045 - 1061 (2014)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Factored Statistical Machine Translation System for English to Tamil Language

Anand Kumar, M.[1]*, Dhanalakshmi, V.[2], Soman, K. P.[1] and Rajendran, S.[1]

[1]*Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, India*
[2]*Department of Tamil, SRM University, Chennai, India*

ABSTRACT

This paper proposes a morphology based Factored Statistical Machine Translation (SMT) system for translating English language sentences into Tamil language sentences. Automatic translation from English into morphologically rich languages like Tamil is a challenging task. Morphologically rich languages need extensive morphological pre-processing before the SMT training to make the source language structurally similar to target language. English and Tamil languages have disparate morphological and syntactical structure. Because of the highly rich morphological nature of the Tamil language, a simple lexical mapping alone does not help for retrieving and mapping all the morpho-syntactic information from the English language sentences. The main objective of this proposed work is to develop a machine translation system from English to Tamil using a novel pre-processing methodology. This pre-processing methodology is used to pre-process the English language sentences according to the Tamil language. These pre-processed sentences are given to the factored Statistical Machine Translation models for training. Finally, the Tamil morphological generator is used for generating a new surface word-form from the output factors of SMT. Experiments are conducted with nine different type of models, which are trained, tuned and tested with the help of general domain corpora and developed linguistic tools. These models are different combinations of developed pre-processing tools with baseline models and factored models and the accuracies are evaluated using the well known evaluation metric BLEU and METOR. In addition, accuracies are also compared with the existing online "Google-Translate" machine translation system. Results show that the proposed method significantly outperforms the other models and the existing system.

ARTICLE INFO

Article history:
Received: 30 May 2013
Accepted: 13 September 2013

E-mail addresses:
m_anandkumar@cb.amrita.edu (Anand Kumar, M.),
dhanagiri@gmail.com (Dhanalakshmi, V.),
kp_soman@cb.amrita.edu (Soman, K. P.),
raj_ushush@yahoo.com (Rajendran, S.)
* Corresponding author

ISSN: 0128-7702

Keywords: Statistical machine translation, preprocessing, English-Tamil machine translation, linguistic tools, morphologically rich language

INTRODUCTION

Machine translation is an automatic translation of one natural language text to another using computer. Now, internet users need a fast automatic translation system between languages. Generally, several approaches such as theLinguistic based and Interlingua based methods are used to develop an automatic machine translation system. Currently, the Statistical Machine Translation (SMT) systems play a major role in developing automatic machine translation between languages. The Statistical Machine Translation method draws the knowledge from an automata theory, artificial intelligence, data structure and statistics. It treats the translation of natural language as a machine learning problem. Learning algorithms produce a model from parallel corpora and using this model, new sentences are translated. Parallel corpora are sentences in one language along with its translation. It is easy to build a bi-lingual baseline SMT system, if sufficient parallel corpora are available. The accuracy of the system is highly dependent on the quality and quantity of the parallel corpus and the domain. The main advantage of using the Statistical Machine Translation is that it is language independent and it disambiguates the sense automatically with the use of large quantity of data. Importantly, SMT systems provide good accuracy for similar language pairs in specific domains or languages that have huge availability of bi-lingual corpora. If the sentences in the language pair are not structurally similar then the translation patterns are difficult to learn by statistical methods. Huge amounts of parallel corpora are required for learning the dissimilar pattern, therefore statistical methods are difficult to use for "less resourced" and dissimilar languages. To enhance the translation performance of dissimilar language pairs and less resourced languages, an external pre-processing is required in the SMT system. Pre-processing includes conversion of source language sentence into similar representation of target language sentence and adding linguistic information using language processing tools.

FACTORED STATISTICAL MACHINE TRANSLATION SYSTEM

The Baseline Statistical Machine Translation system only considers the surface word-forms of sentences and does not include the linguistic knowledge of the languages, therefore its performance is significantly less for dissimilar language pair when compared to similar language pair. To resolve this issue, factored models are introduced in SMT system. The factored model, which is a subtype of phrase based SMT (Philipp Koehn & Hieu Hoang, 2007), will allow multiple levels of representation of the word from the most specific level to more general levels of analysis such as lemma, part-of-speech and morphological features. The phrase based translation model is based on the noisy channel models. Bayes

rule is used to reformulate the translation probability for translating a source language sentence into target language sentence. The objective of the translation model is to find the probability of target language sentence *T*, given a source language sentence *S*.

$$P(T/S) = (P(S/T)P\ (T)) \ / \ P(S) \quad [1]$$

$$\check{T} = argmax\ P(S/T)P\ (T)/P(S) \quad [2]$$

$$\check{T} = argmax\ P(S/T)P\ (T) \quad [3]$$

From the equation (2), the denominator *P(S)* is removed, since the probability of the source sentence is constant. *P(S/T)* is given by translation model and *P(T)* is given by language model. In addition, to find a best translation a decoder is required, which given a source sentence *S*, produces the best probable target sentence *T*, or possibly an n-best list of the most probable translations. The probability of best translation is calculated from the translation probability and language model and *argmax* chooses the highest probable one (*T*) among the all possible target language sentences (*T*). Factored translation models can be seen as the combination of several components (language model, reordering model, translation steps, and generation steps) (Philipp Koehn & Hieu Hoang, 2007). These components define one or more feature functions that are combined in a log-linear model.

$$P\left(\frac{t}{s}\right) = \frac{1}{Z} exp \sum_{i=1}^{n} \lambda_i \ h_i\ (t, s) \quad [4]$$

z is a normalization constant that is ignored in practice. Evaluate each feature function h_i to compute the probability of a translation *t* given an input sentence *s*.

RELATED WORKS

This section discusses the literature review about adding linguistic information into the Statistical Machine Translation system and existing English to Tamil Machine Translation systems.

Reordering methods using linguistic knowledge attained a significant improvement in performance for translation from French to English (Xia & Mc-Cord, 2004) and from German to English (Collins *et. al*., 2005). Panagiotis (2005) proposed a novel algorithm for incorporating morphological knowledge from English to the Greek Statistical Machine Translation (SMT) system. She/He suggested a method for improving the translation quality of existing SMT systems, by incorporating word-stems. Avramidis *et. al*., (2008) addressed the problem of translating morphologically poor language into morphologically rich language and the improvement in performance is shown for translating from English to Greek and English to Czech. Ananthakrishnan R *et. al*., (2008) developed a syntactic and morphological pre-processing for the English to Hindi SMT system. They reordered the English source sentence as per Hindi syntax, and segmented the suffixes of Hindi for morphological processing. Sara Stymne (2009) explored how compound processing can be used to improve the accuracy of Phrase-Based

Statistical Machine Translation (PBSMT) between English and German/Swedish. For translation into Swedish and German the segmented parts are merged after translation. Ann Clifton (2010) examined various pre-processing methods for augmenting SMT models with morphological information to improve the quality of English-Finnish automatic translation task. Reyyan Y *et. al*., (2010) reported a novel scheme for translating languages with very disparate structure. In this method, syntax of the source language sentence is mapped to morphology of the target language sentence in Factored Statistical Machine Translation.

Various automatic machine translation systems have been developed for translating the English language to the Tamil language. Ulrich Germann (2001) reported his experience with building a statistical MT system from scratch, including the creation of a small parallel Tamil-English corpus. Fredric C.Gey (2002) reported the prospects of machine translation of the Tamil language. The major problems in connection with machine translation and cross-language retrieval of Tamil (and other Indian languages) are discussed. Vasu Renganathan (2002) proposed an interactive approach to develop a web based English to Tamil machine translation system. AUKBC research centre developed a Human Aided Machine Translation System from English to the Tamil language. This machine translation system has three major components, viz. source language morphological analyzer, mapping unit and the target language generator. This prototype version handles simple sentences and only works for limited vocabulary and grammar. Vetrivel *et. al*., (2010) proposed a statistical based machine translation system using HMM based alignment for words and phrases in a parallel text and Tamil transformation rules and word combination rules are also used. Loganathan R (2010) developed the English-Tamil machine translation system using rule-based and corpus-based approaches. For the rule based approach, the structural difference between English and Tamil is considered and syntax transfer based methodology is adopted for translation. Saravanan *et. al*., (2010) developed a Rule based Machine translation system for English to Tamil. Saraswathi *et. al.,* (2011) developed a machine system for English to Tamil as well as Tamil to English using rule based Machine Translation and knowledge based Machine Translation. Loganathan R (2012) developed the English-Tamil statistical machine translation system using morphological processing. He separated the morphological suffixes of English and Tamil to improve the quality of phrase based and hierarchical machine translation systems. Using the statistical machine translation approach, Google developed a web based machine translation engine for English to Tamil language. The phrase and word selection is excellent in this system but it failed to produce morphologically fluent Tamil sentences for even simple English sentences. Even though machine translation research is started in the 1950s, however, high-quality English-Tamil Machine translation system is not available

at present. Statistical models require huge amounts of parallel data, which are not readily available for English-Tamil pair.

METHODS

Overall System Architecture

Automatic machine translation into morphologically-rich languages remains a highly challenging task because manual translation itself is difficult. Tamil is a morphologically rich language with free word-order and English is a morphologically simple language with the fixed word order. This morphological and structural divergence increases the challenges in translating from English to Tamil language. The overall architecture of the proposed English to Tamil factored SMT system is illustrated in Fig.1. In this figure, the training of the SMT system is shown using dotted and bold lines. The dotted line represents the formation of the language model and the bold line denotes the creation of translation model. The light blue line shows the testing of the SMT system.

The pre-processing module is externally attached to the SMT system. This module converts the bilingual corpora into factored bilingual corpora using morphology based linguistic tools and reordering rules. After pre-processing, the representation of the source language syntax is closely follows the structure of the target language. This transformation decreases the complexity in alignment, which is a key problem in baseline SMT system. Parallel corpora and monolingual corpora are used to train the statistical translation models. Parallel corpora are collected and converted into factored parallel corpora using pre-processing. English sentences are factored using Stanford Parser and Tamil sentences are factored using Tamil POS Tagger and Morphological analyzer. The Monolingual corpus is collected from various online newspaper websites and then used in the Language model. Finally, in post processing, the Tamil morphological generator is used for generating Tamil surface words from the output factors of SMT decoder.

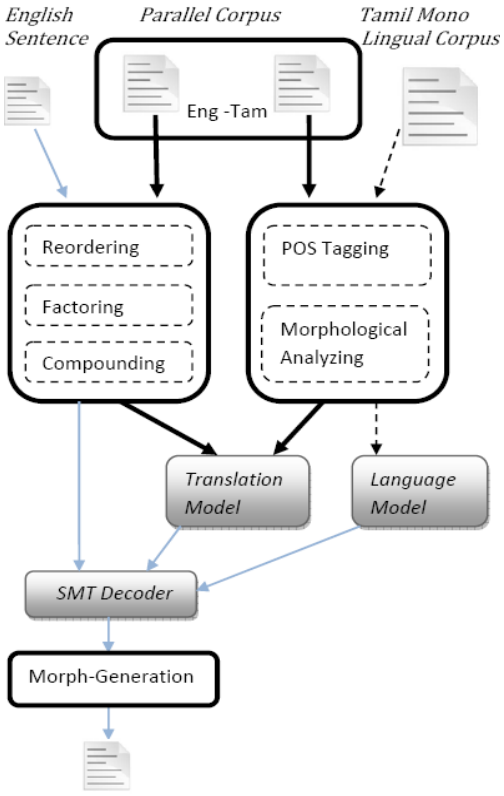

Fig.1: The Factored SMT system for English to Tamil language

Details of Pre-processing English Language Sentences

The proposed preprocessing module for the English language sentence consists of reordering, factorization and compounding. This language specific pre-processing prior to translation notably improves the translation quality.

Reordering English Language Sentences

Reordering means, rearranging the word order of one natural language sentence into the word order that is closer to that of another natural language sentence. It is an important task in translation for languages which differs in their syntactic structure. English and the Tamil language pair has disparate syntactic structure. The word order of the English language sentence is Subject-Verb-Object (SVO) whereas in Tamil sentence, the word order is Subject-Object-Verb (SOV). For instance, the main verb of a Tamil sentence always comes at the end but in English it comes between the subject and object. English syntactic relations are retrieved from the Stanford Parser tool (Klein & D Manning, 2003). Based on the developed reordering rules, the source language sentence is reordered. Reordering rules are handcrafted based on the syntactic word order difference between English and the Tamil language. One hundred and eighty reordering rules are created based on the structure of English and Tamil. Sample reordering rules are shown in Table 1 Reordering significantly improves the performance of machine translation system. Automatic Lexicalized reordering is implemented in the Moses toolkit. Automatic reordering in this toolkit is not a language specific method so it is not good for short range and simple sentences. Therefore, the external component is needed for dealing with the sentences which are not reordered properly.

TABLE 1
Reordering rules for English language sentences

Source	Target
S -> NP VP	# S -> NP VP
PP -> TO NP-PRP	# PP -> TO NP-PRP
VP -> VB NP* SBAR	# VP -> NP* VB SBAR
VP -> VBD NP	# VP -> NP VBD
VP -> VBD NP-TMP	# VP -> NP-TMP VBD
VP -> VBP PP	# VP -> PP VBP

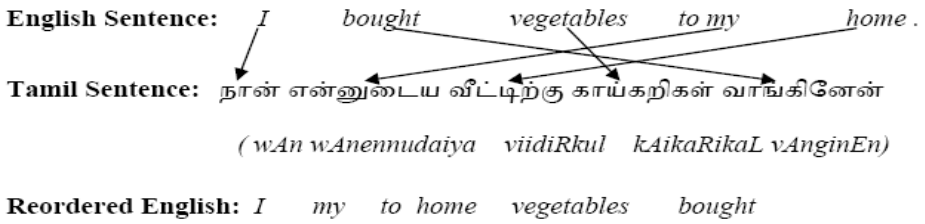

Fig.2: Reordering of an English language sentence

Factorization of the English Language Sentence

Factored models are predominantly used for morphologically rich languages, in order to reduce the amount of bilingual data. Factorization refers to splitting the word into linguistic factors and integrating it as a vector. The English parsed tree which is obtained from Stanford Parser is used to retrieve the linguistic information such as lemma, part-of-speech tags, syntactic information and dependency information. This linguistic information is integrated as factors in the original English word. Factorization is one way of representing morphological knowledge to Statistical machine translation explicitly. Factorization also reduces the Out-Of-Vocabulary (OOV) rate. Table 2 shows the factors of words in an example of an English language sentence. In this table, *word* refers surface word, *lemma* represents the dictionary word or root word, *w-c* represents word-class category and *morph* tag represents compound tag which contains morphological information and/or function words. In some cases the *morph* tag also contains the dependency relations and/or PNG (Person-Number-Gender) information.

Compounding for English Language Sentence

During automatic translation from morphologically simpler language to morphological rich language, it is very hard to retrieve the required morphological information from the source language sentence. This morphological information is an important term for producing an exact target language word-form. Morphologically rich languages have a large number of surface forms in the lexicon to compensate for a free word-order. This large number of word-forms in the Tamil language is very difficult to generate automatically from the English language words. The pre-processing phase compounding is referred as adding extra morphological information to the morphological factor of the source (English) language words. Additional morphological information includes function words, subject information, dependency relations, auxiliary verbs, and model verbs. This information is based on the morphological structure of target language sentence. In compounding, English function words are identified from the factored corpora using dependency information and these identified function words are included in a morphological factor

TABLE 2
Factors of Words in English sentence

WORD	LEMMA	W-C	MORPH	FACTORS
I	I	PRP	PRP	I | i | PN | prn
bought	buy	V	VBD	bought | buy | V | VBD
vegetables	vegetable	V	NNS	vegetables | vegetable | N |NNS
to	to	PRE	TO	to | to | TO | TO
my	my	PRP	PRP$	my | my | PN | PRP$
home	home	N	NN	home | home | N | NN

of the corresponding content word. Finally these function words are removed from the factored sentence.

For instance, the sentence "*I bought vegetables to my home*" is pre-processed. The word "*to*" is identified as a function word and it is removed from the sentence and attached to the morphological factor of the word "*home*". The main reason to perform the compounding process in English sentence is that the words like "*to*" (or any prepositions) does not has the equivalent individual word in Tamil. Actually, "***to*** *home*" in English sentence is equivalent to the Tamil word "*vitti**Rkku***". Compounding reduces the length of the English language sentence during pre-processing. Similar to the function words, auxiliary verbs and model verbs are also removed from sentence and attached in a morphological factor of the corresponding content word or head word. Now the representation of the English language sentence is similar to that of the Tamil language sentence. This compounding step indirectly integrates dependency information and other required morphological information into the source language factor.

Details of Pre-processing Tamil Language Sentence

Similar to the pre-processing of English language sentence, the Tamil language sentences are also pre-processed using linguistic tools such as POS tagger and morphological analyzer. Tamil surface words are segmented into linguistic units and these segments are annotated and integrated as linguistic factors in SMT training corpora. At first, the Tamil sentence is given to the Tamil Part-of-Speech Tagger tool (Dhanalakshmi V *et. al*., 2008) and then using the part-of-speech information, the minimized part-of-speech tag (or Course-grained tag) is identified. Based on the minimized tag, the words are given to the Tamil morphological analyzer tool (Anand Kumar *et. al*., 2010b). The Morphological analyzer splits the word into lemma and morphological information. Pre-processing is carried out in parallel corpora as well as the monolingual corpora. The pre-processing phase in the Tamil language converts the corpora into factored corpora. Tamil words and its factors are shown in Table 3. Fig.3 and Fig.4 show the alignment of the English and Tamil sentences before and after pre-processing.

TABLE 3
Tamil factored sentence

WORD	FACTORS
நான்	நான் | நான் |P| null
என்னுடைய	என்னுடைய|என்|P| poss
வீட்டிற்கு	வீட்டிற்கு|வீடு|N| DAT
காய்கறிகள்	காய்கறிகள்|காய்கறி|N|PL
வாங்கினேன்	வாங்கினேன்|வாங்கு|V|PAST_1S

Factored SMT System for English to Tamil Language

The Statistical Machine Translation consists of three key components viz. translation modeling, language modeling and decoding. These components are implemented using GIZA++, SRILM and Moses toolkits. GIZA++ is a statistical machine translation toolkit that is used to train IBM models 1-5 and an HMM word alignment model. It is an extension of GIZA which is designed as a part of the SMT toolkit. SRILM is a toolkit for language modeling that is used in speech recognition, statistical tagging and statistical machine translation. Moses is an open source statistical machine translation toolkit that allows to automatically training the translation models for any language pair. A collection of parallel translated texts is only required for a language pair. An efficient search algorithm finds quickly the highest probability translation among the exponential number of choices. Morphologic, syntactic and semantic information are integrated in preprocessing. Pre-processed English and Tamil language sentences are used in SMT training. Fig.5 explains the mapping of English factors and Tamil factors in Factored SMT System.

Initially, English factors "Lemma" and "Minimized-POS" are mapped into the Tamil factors "Lemma" and "Minimized-POS Tag" then "Minimized-POS" and "Compound-Tag" factors of English language is mapped to "Morphological information" factor of Tamil language.

Here, the remarkable thing is that the Tamil surface word forms are not

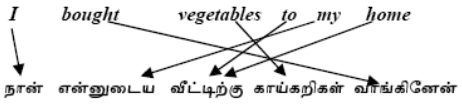

Fig.3: Alignment before pre-processing

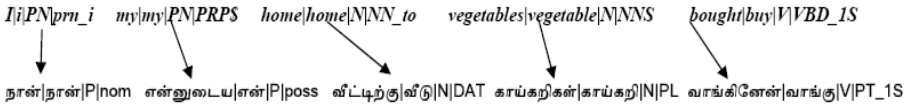

Fig.4: Alignment after pre-processing

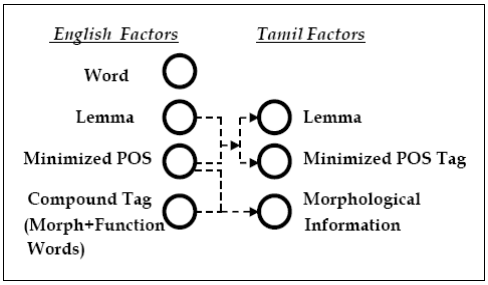

Fig.5: Mapping English factors to Tamil Factors

generated using the SMT decoder, only the factors are generated from SMT and the new surface word is generated in the post-processing stage. This is because the target language Tamil is morphologically rich and the parallel corpora which is used in this system is also small in size. The Tamil morphological generator is used in post-processing to generate a Tamil surface word from output factors. The developed English-Tamil prototype machine translation system properly handles the noun-verb agreement. This is an essential requirement for translating into morphologically rich languages like Tamil.

Post Processing for English to Tamil Factored SMT

Post processing is applied to generate a Tamil surface word from the output factors. In this proposed factored SMT system, the main aim is to translate the factors only, not to generate the surface word-form. Due to the morphological rich nature of Tamil language, word generation is handled separately.

Tamil Morphological Generator

The Morphological generator is a language processing tool which is used to generate a surface word from its lemma and morphological description. It is a reverse process of morphological analyzer. Tamil morphological generator (Anand Kumar *et. al*., 2010a) receives the output factors from the SMT decoder in the form of "*lemma* + *word_class* + *morpho-lexical information*", where *lemma* denotes the lemma of the word form to be generated, *word_class* specifies the grammatical category and *morpho-lexical information* specifies the type of inflection.

EXPERIMENTS AND RESULTS

This section describes the experimental setup and data used in the English to theTamil statistical machine translation system. The corpora consist of approximately 8.5K English to Tamil parallel sentences. General domain English-Tamil parallel corpora are used in the experiments. The training set is built with 6,500 parallel sentences and a test set is constructed with 1462 sentences. 500 parallel sentences are used for tuning the system. For language model, sizes of 90k Tamil sentences are used. Total words and average word length of sentences in baseline and pre-processed parallel corpora used in these experiments are shown in the Table 4 and 5. After pre-processing the average word length of the English sentences are reduced, according to the word-length in Tamil sentences.

Nine different types of models are trained, tuned and tested with the help of parallel corpora. The general categories of the models are Baseline and Factored systems. The detailed models are,

1. Baseline (BL)
2. Baseline with Automatic Reordering (BL+AR)
3. Baseline with Rule based Reordering (BL+RR)
4. Factored system + Morph-Generator

(Fact)

5. Factored system + Auto Reordering +Morph-Generator (Fact+AR)
6. Factored system +Rule based Reordering + Morph-Gen (Fact+RR)
7. Factored system + Compounding + Morph-Generator (Fact+Comp)
8. Factored system + Auto Reordering +Compounding +Morph-Generator (Fact+AR+Comp)
9. Factored system +Rule based Reordering +Compounding+ Morph-Generator (Fact+RR+Comp)

For a baseline (BL) system, a standard phrase based system is built using the surface forms of the words without any additional linguistic knowledge and with a 4-gram language model in the decoder. Cleaned and tokenized raw parallel corpus is used for training the system. Lexicalized reordering model (msd-bidirectional-fe) is used in the baseline with automatic reordering (BL+AR) model. Another baseline system is built with the use of rule based reordering (BL+RR). In all the developed factored models, the Tamil morphological generator is commonly used in post processing stage.

Instead of using the surface form of the word, a root, part-of-speech and morphological information are included into the word as an additional factors in factored machine translation system. A factored parallel corpus is used for training the system. English factorization is done by using Stanford Parser tool and for Tamil, POS Tagger (Dhanalakshmi V *et. al*., 2008) and Morphological analyzers (Anand Kumar *et. al*., 2010b) are used to factor the sentence. In this factored model, a token/word is represented with four factors as *Surface|Root|Wordclass|Morpho logy*. The first factored model (Fact) is built without Reordering and Compounding the English sentences. Factored system with

TABLE 4
Details of Baseline Parallel corpora

Corpora		Total Sentences	Total Words		Average Word Length	
			English	Tamil	English	Tamil
General	Training	6500	56760	34926	8.732	5.3723
	Tuning	500	4144	2684	8.288	5.368
	Testing	1462	8860	-	6.060	-

TABLE 5
Details of Pre-processed Parallel corpora

Corpora		Total Sentences	Total Words		Average Word Length	
			English	Tamil	English	Tamil
General	Training	6500	45317	34926	6.97	5.3723
	Tuning	500	3405	2684	6.81	5.368
	Testing	1462	6554	-	4.482	-

lexicalized reordering (Fact+AR) and rule based reordering (Fact+RR) models are also constructed to discover the impact of reordering in the performance of the Factored statistical machine translation system. Another factored system is built with the use of Compounding (Fact+Comp). Here the *Morphology* factor contains morphological information and function words on English side, and morphological tags on Tamil side. Factored system with Compounding is developed with lexicalized reordering (Fact+AR+Comp) and rule based reordering (Fact+RR+Comp). In this model, English words are factored and reordered using the developed rules. In addition to this Compounding is also performed in the English language side. The pre-processing methodology used in this paper reduces the Out-Of-Vocabulary (OOV) words drastically. Table 6 shows the number of OOV words and OOV rate for the developed models. The approximate OOV rate for all the Baseline models are 0.24 and for Factored models, it ranges from 0.136 to 0.254. In factored models, compared to the other systems, Compounding based models provide high OOV rate except the final one. Whereas, rule based reordering reduces the OOV rate in all the models.

All the developed models are evaluated with the same test-set which contains 1462 English sentences. The well known Machine Translation metrics BLEU (Kishore Papineni, 2002) and METOR (Alon Lavie, 2010) are used to evaluate the developed models. In addition to that the existing "Google Translate" online English-Tamil machine translation system is also evaluated to compare with the developed models. The results are in terms of Mert-BLEU, Multi-BLEU and METOR score and it is shown in the Table 7. Table 8 indicates the Lemma-wise scores for the developed factored models. Mert-BLEU represents the Minimum Error Rate Tuning BLEU score which is obtained while tuning. Multi-BLEU perl script in Moses toolkit is used for evaluating the multi-BLEU scores. Table 7 depicts the Baseline and Factored models' performance in terms of a well-known machine translation metrics BLEU

TABLE 6
Out-of-Vocabulary Rate

	Models	Number of OOV Words	OOV Rate
BASELINE	BL	2134	0.240
	BL+AR	2134	0.240
	BL+RR	2142	0.241
FACTORED	Fact+Mgen	1617	0.182
	Fact+AR+Mgen	1617	0.182
	Fact+RR+Mgen	1205	0.136
	Fact+Comp+Mgen	2256	0.254
	Fact+AR+Comp+Mgen	2256	0.254
	Fact+RR+Comp+Mgen	1104	0.168

and Metor. Here, the scores are calculated by considering the word's surface level (word-from). The proposed method shows that 18% Multi-Bleu score improvement and 61% Metor score improvement against the existing "Google Translate" system.

Table 8 shows the lemma-wise accuracies of Factored models. In factored models, lemma with word-class information in English is getting translated into lemma and word-class in Tamil. So the lemma-wise accuracies are calculated in the factored model only. Similar to word-wise accuracies, lemma-wise accuracies for the proposed system (Fact+RR+Comp+Mgen) shows the improvement in BLEU and METOR evaluation scores.

In addition to that, the developed F-SMT based system also handles the noun-verb agreement perfectly. This is an important and challenging job for translating into morphologically rich languages like Tamil. The example given below shows the comparison between the proposed system and existing system in-terms of agreement handling. The compounding phase in the proposed system maps the English dependency and subject information into the Tamil morphology and this mapping handles the noun verb agreement accurately.

TABLE 7
BLEU and Metor Score

	Models	Mert-BLEU	Multi-BLEU	Metor
BASELINE	BL	0.0107	1.13	0.123
	BL+AR	0.0112	0.92	0.121
	BL+RR	0.0098	0.85	0.121
FACTORED	Fact+Mgen	0.035	1.18	0.138
	Fact+AR+Mgen	0.0343	0.81	0.075
	Fact+RR+Mgen	0.0368	0.82	0.126
	Fact+Comp+ Mgen	0.0335	1.55	0.123
	Fact+AR+Comp+Mgen	0.0337	1.39	0.123
	Fact+RR+Comp+Mgen	0.0414	7.86	0.377
	Google Translate	-	6.66	0.234

TABLE 8
BLEU and Metor Scores for Lemma

Models	Multi-BLEU	BLEU-1	BLEU-4	Metor
Fact+Mgen	15.50	49.9	4.0	0.377
Fact+AR+Mgen	12.25	49.5	2.0	0.366
Fact+RR+Mgen	18.70	53.2	5.8	0.418
Fact+Comp+Mgen	5.19	38.9	0.9	0.217
Fact+AR+Comp+Mgen	5.33	39.0	0.9	0.217
Fact+RR+Comp+Mgen	30.23	57.7	16.6	0.499

English Sentence:
I went to school with her.

Google Translate's Output:
நான் அவளுடன் பள்ளி
சென்றார்

Proposed System's Output:
நான் அவளுடன் பள்ளிக்கு
சென்றேன்.

CONCLUSION AND FUTURE WORK

Automatic machine translation is a challenging task for languages which are different in morphological structure and word order. For training the SMT system, both monolingual and bilingual sentence-aligned parallel corpora of significant size are essential. But, in most of the cases only small amounts of bilingual corpora are available for the desired domain and language pair. Therefore, linguistic knowledge is used in SMT to reduce the need of massive amounts of data. This is especially desirable for the English-Tamil language pair where massive amounts of parallel corpora are not available.

This paper presents the novel methods for incorporating linguistic knowledge in SMT to achieve an enhancement in the English to Tamil statistical machine translation system. Most of the techniques proposed in this paper can be applied directly to other language pairs especially for translating from morphologically simple language to morphologically rich language. The precision of the translation system depends on the performance of each and every module and the language processing tools used in the system. The experimental results clearly demonstrate that the new techniques proposed in this paper are definitely significant. The precision of the developed system is also affected by the accuracy of the Tamil morphological generator system, totally nine different SMT models are experimented with general domain corpora and the BLEU and METOR scores are compared. The existing "Google Translate" system is also evaluated and compared in our experiments. This proposed system produces morphologically fluent Tamil sentences for the given English sentences whereas, the existing Google translate system failed to produce it. Except for the system developed by Loganathan (2012), all the other English-Tamil machine translation systems are not evaluated by the well known metrics. The experiments are also performed with automatic lexical reordering on the source sentence. The developed model (Factored SMT with Reordering and Compounding) has reported a 0.377 METOR and 7.86 BLEU score for English to Tamil translations. Adding pre and post processing in factored SMT provided considerable improvement in BLEU over a phrase based baseline system and the factored baseline system. Improvement in BLEU and METOR evaluation scores shows that this proposed method is appropriate for developing SMT system from morphologically simple language to morphologically rich languages.

In future, adding more training data with different sentence structures definitely will improve the accuracy of the proposed system. Additionally, incorporating modules

for dealing idioms and phrases and splitting of complex sentences will also enhance the performance of the proposed system. Finally the conclusion is that morphologically rich languages need extensive morphological pre-processing before the SMT training to make the source language structurally similar to target language and it also needs an efficient post-processing in order to generate the surface word correctly.

REFERENCES

Abeera, V. P., Aparna, S., Rekha, R. U., Anand Kumar, M., V. Dhanalakshmi, V., Soman, K. P., & Rajendran, S. (2012).Morphological Analyzer for Malayalam Using Machine Learning, *Data Engineering and Management*, Springer Heidelberg publishers, 253-254, DOI: 10.1007/978-3-642-27872-3_38.

Alon Lavie & Michael Denkowski. (2010). The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*.

Anand Kumar, M., Dhanalakshmi, V., Rekha, R. U., Soman, K. P., & Rajendran, S. (2010a). A Novel Data Driven Algorithm for Tamil Morphological Generator*, International Journal of Computer Applications (IJCA)* 6(12):52–56, September 2010. DOI:10.5120/1121-1470

Anand Kumar, M., Dhanalakshmi, V., Soman, K. P., & Rajendran, S. (2010b). A Sequence Labeling Approach to Morphological Analyzer for Tamil Language, *International Journal on Computer Science and Engineering (IJCSE)* Vol. 02, No. 06, 2010, 2201-2208.

Ananthakrishnan, R., Pushpak Bhattacharya, Jayprasad Hegde, Ritesh M.Shah, & Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. *In IJCNLP 2008: Third International Joint Conference on Natural Language Processing*, Hyderabad, India. Rochester, NY.

Antony, P. J., Anand Kumar, & K. P. Soman, (2010), Paradigm based Morphological Analyzer for Kannada Language Using Machine Learning Approach, *International journal on Advances in Computational Sciences and Technology*, ISSN 0973-6107 Volume 3 Number 4 (2010) pp. 457–481.

Clifton, Ann. (2010). Unsupervised Morphological Segmentation For Statistical Machine Translation. *Master of Science thesis*, Simon Fraser University.

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC* (Vol. 6, pp. 449-454).

Dhanalakshmi, V., Anand Kumar, M., Vijaya, M. S., Loganathan, R., Soman, K. P., & Rajendran, S., (2008). Tamil Part-of-Speech tagger based on SVMTool, *In Proceedings of the COLIPS International Conference on natural language processing(IALP)*, Chiang Mai, Thailand.

Eleftherios, A., & Philipp, K. (2008). Enriching morphologically poor languages for statistical machine translation, *In Proceedings of ACL-08/ HLT*, pages 763–770, Columbus, Ohio.

Fei Xia, Michael,. McCord. (2004). Improving a statistical MT system with automatically learned rewrite patterns. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 508–514, Geneva, Switzerland. Association for Computational Linguistics.

Fredric C. Gey , (2002), Prospects for Machine Translation of the Tamil Language, *Proceedings of international Tamil Internet conference 2002*.

George Doddington (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proceeding of the ARPA*

Workshop on Human Language Technology.

Germann, U. (2001). Building a statistical machine translation system from scratch: how much bang for the buck can we expect?, *In Proceedings of the workshop on Data-driven methods in machine translation*, 1–8, ACL, Morristown, NJ, USA.

Jes´us Gim´enez & Llu´ıs M`arquez. (2004), "SVMTool: A general POS tagger generator based on support vector machines", *Proceedings of the 4th LREC Conference*.

Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th Meeting of the Association for Computational Linguistics* (ACL'02) (pp. 311–318). Philadelphia, PA.

Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*-Volume 1 (pp. 423-430). Association for Computational Linguistics.

Koehn, P., & Hoang, H. (2007). Factored translation models. *In Proceedings of the EMNLP 2007*, Prague, Czech Republic.

Loganathan, R. (2010). English-Tamil Machine Translation System. *Master of Science by Research Thesis*, Amrita Vishwa Vidyapeetham, Coimbatore.

LoganathanRamasamy OndrejBojar, Zdenek Žabokrtský, L. O. (2012), Morphological Processing for English-Tamil Statistical Machine Translation. In 24th *International Conference on Computational Linguistics* (p. 113).

Marta Ruiz Costa-jussa & J. A. R. Fonollosa (2009), State-of-the-art word reordering approaches in statistical machine translation, *IEICE Transactions on Information and Systems*, vol. 92, no. 11, pp. 2179–2185, November 2009.

Michael Collins., Philipp Koehn, & Ivona Kuˇcerová (2005). Clause restructuring for statistical machine translation. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 531–540, June 2005.

P. Karageorgakis, P., Potamianos, A., & Klasinas, I. (2005). "Towards incorporating language morphology into statistical machine translation systems," *in Proc. Automatic Speech Recogn. and Underst. Workshop.*

Rabih M. Zbib. (2010). Using Linguistic Knowledge in Statistical Machine Translation, *Ph.D. thesis*, Massachusetts Institute Of Technology, 2010.

Rajendran, S., Viswanathan, S., & Ramesh Kumar. (2003). Computational morphology of verbal complex, *Language in India*, Volume 3:4.

Reyyan Yeniterzi & Kemal Oflazer. (2010). Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish , *Proceedings of the 48th Annual Meeting of the ACL*, pages 454–464, Uppsala, Sweden, 11-16 July 2010.

Sai Kiranmai, G., Mallika, K., Anand Kumar, M., Dhanalakshmi, V., & Soman, K.P. (2010). Morphological Analyzer for Telugu Using Support Vector Machine, *Information and Communication Technologies*, Springer Berlin Heidelberg, ISBN: 978-3-642-15766-0, pp430-433.

Sara Stymne. (2009) Compound Processing for Phrase-Based Statistical Machine Translation. *Licentiate thesis*, Linköping University, Sweden.

Saraswathi,S., Kanivadhana, P., Anusiya, M., & Sathiya, S. (2011). Bilingual Translation System (For English and Tamil), *International Journal on Computer Science and Engineering (IJCSE)*, *3*(3).

Saravanan, S., Menon, A. G., &Soman, K. P. (2010). English to Tamil Machine Translation System, *Proceedings of INFITT-2010*, at Coimbatore.

Soha sulthan. (2011). Applying morphology to English Arabic statistical Machine Translation. *Master of Science thesis*, Eidgenössische Technische Hochschule Zürich

Vasu Renganathan. (2002). An ineractive approach to development of English to Tamil machine translation system on the web. *Proceedings of INFITT-2002*.

Vetrivel, S. & Baby, D. (2010). English to Tamil statistical machine translation and alignment using HMM. *Proceedings of the 12th international Conference on Networking, VLSI and Signal Processing*, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 182-186.