Syntactic Based Machine Translation from English to Malayalam

Anitha T Nair
Department of Computer Science
Federal Institute of Science and Technology
Angamaly, Ernakulam
anitha.mrt@gmail.com

Sumam Mary Idicula
Department of Computer Science
Cochin University of Science and Technology Cochin, India
Cochin, India
sumam@cusat.ac.in

Abstract— **Due to the emergence of multiple language support on the Internet, machine translation (MT) technologies are indispensable to the communication between speakers using different languages. Recent research works have started to explore tree-based machine translation systems with syntactical and morphological information. This work aims the development of Syntactic Based Machine Translation from English to Malayalam by adding different case information during translation. The system identifies general rules for various sentence patterns in English. These rules are generated using the Parts Of Speech (POS) tag information of the texts. Word Reordering based on the Syntax Tree is used to improve the translation quality of the system. The system used Bilingual English –Malayalam dictionary for translation.**

Keywords-morphology;POS;SBMT;

I. INTRODUCTION

Machine translation (MT) is the study of designing systems that translates text from one natural language to another [11]. It is an interesting and challenging field of Computer Science This is one of the most important application areas of Natural Language Processing (NLP). Basic task of translation is undertaken by a computer program in conjunction with automated dictionaries and grammars. Ability of adjustment to context is the difficulty associated with machine translation. Due to the lack of proficiency in English, information access is very difficult for the common people. This inaccessibility problem can be avoided by the use of MT system. The advantage of MT is the speed at which it translates the text. The MT system is usually integrated with various NLP activities such as information retrieval and summarization.

Earlier versions of MT system are rule based. Various techniques used in Rule Based Machine Translation (RBMT) systems are direct translation, interlingua and transfer based [2].Direct MT system directly translate the source language (SL)to the target language (TL) in which words in the SL is replaced with the words in the target language. In direct MT system, we have to create a new system for every new language pair. Interlingua representation converts the input to a common internal representation. This representation is common to all languages used by the system. It is not accurate for certain translation systems .As a consequences, it became widely accepted that the less ambiguous transfer approach offered better prospects. In 1989 new methods and strategies for MT system were introduced. These recent approaches are Corpus Based, Knowledge based and hybrid approaches. Statistical Machine Translation (SMT) [12] and Example Based Machine Translation (EBMT) [15] come under corpus based approach. Corpus based methods require very large volumes of parallel texts for the source and target languages. This transfer based system transform SL to a form that depends only on the target language. The output text is generated from this form. This approach is less ambiguous.

In this paper a Syntactic Based Machine Translation (SBMT) for English to Malayalam is presented. This is basically a transfer based bilingual system. Bilingual MT systems can be specifically designed for two particular languages. It is based on an idea of re-arranging nodes of parse tree of the source language to the target language [10]. General rules are identified for particular types of sentence patterns and this pattern is used for translation. Bilingual English-Malayalam dictionary and a morphology generator are used for translation.

The rest of the paper is organized as follows .Section II covers some related work in the MT area .System overview is described in section III. The performance of the SBMT system is evaluated in section IV. The paper is concluded in section V.

II. RELATED WORK

Historically different approaches to MT have been used. MT for languages with limited amount of resources is a challenging task [2]. New rules for Rule based MT are presented in [13].A study of example based MT is a reported in [15]. Phrase based statistical machine translation is discussed in [4].It uses the context free grammar for translation. A pattern directed rule based system is discussed in ANGALABARATHI [1]. It is a machine aided translation methodology specifically designed for translating English to all Indian languages. AngalaHindi [3] is a English to Hindi version of ANGALABARATHI. Besides using rule base it also uses an example base for translating frequently used noun phrases and verb phrases.

One of the main research areas in MT is phrase reordering models. Agreement and word reordering problems in English to Arabic MT is examined in [13]. Subject Verb Object (SVO) reordering can be done in different ways .One such method is based on POS tags. Word reordering in SMT with a POS based distortion model is described in [7]. Principles of Discriminative Reordering Models for Statistical Machine Translation is presented in [5].Another reordering method is

978-1-4673-2149-5/12/$31.00 ©2012 IEEE

the syntax based reordering. Syntactic reordering within English to Arabic translation task investigates in [10]. SBMT system described in this paper uses syntax based reordering.

III. SYSTEM OVERVIEW

The architecture of Syntactic Based Machine translation (SBMT) is given in Figure 1.It is specifically designed for translating text in English to Malayalam. English Text is the input to this system. For the translation purpose this system uses a bilingual English-Malayalam dictionary and a morphology generator [6]. After performing necessary transformation Malayalam text is generated as the output. This system mainly consists of four modules. They are

i. Syntax tree generator: Preprocessing of the input text is carried out in this module.

ii Word Reordering: Converts SVO form to SOV.

iii. Pattern Recognition: Identify the sentence pattern of the text

iv. Translation: Translate the reordered English text to Malayalam text.

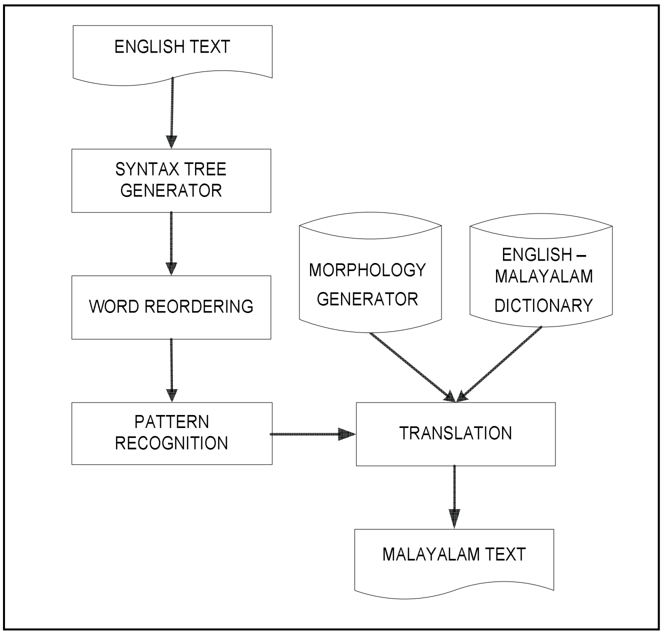

Figure 1. System architecture

A. Syntax Tree Generator

In this sub system, source sentence is given to the Stanford Parser. The parser examines the syntactic structure of the sentence. Using the parser the syntax tree, Parts of speech (POS) tag and dependency informations are generated. Part-of-speech tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adverb, adjective or other lexical class marker to each word in a sentence. Dependency information represents dependencies between individual words.

Example:

English Text: He bought a watch for his wife.

Output generated by the parser is shown in the table Table I

TABLE I

OUTPUT INFORMATION

Parse Tree	(ROOT (S (NP (PRP He)) (VP (VBD bought) (NP (DT a) (NN watch)) (P (IN for) (NP (PRP$ his) (NN wife)))) (. .)))
POS Tag:	He bought a watch for his wife PRP VBD DT NN IN PRP$ NN
Dependency Information	nsubj (bought-2, He-1) det (watch-4, a-3) dobj (bought-2, watch-4) prep (bought-2, for-5) poss (wife-7, his-6) pobj (for-5, wife-7)

B. Word Reordering

In machine translation (MT), one of the main problems to handle is word reordering [5]. Different languages differ in their syntactic structure. English sentences normally follows Subject, Verb and Object format whereas Malayalam sentences follows Subject, Object and Verb format. The main verb is always at the right end of the Malayalam sentence. This difference gives rise to the need of rearranging the structure of source language with respect to the target language. Most of the reordering techniques are purely statistical processes and no syntactic knowledge of the language is used for processing [7]. Syntactical reordering improves the translation quality of machine translation systems. Many reordering models have recently been used in different MT systems.

This approach follows a new method for syntax based reordering. In order to reorder, different phrases are extracted from the syntax tree. Phrases are considered as the second level of classification as they tend to be larger than individual words, but are smaller than sentences Different phrases constitute a sentence. According to 21 basic conversion rules, these phrases are grouped as per the word group of Malayalam. Order conversion rule are described in Table II.

Example:

English Text: He bought a watch for his wife.

Phrases:

NP	He
NP	a watch
NP	his wife
PP	for his wife
VP	bought a watch for his wife

TABLE II

REORDERING RULES

English	Malayalam
NP->NP PP	PP NP
NP->NP VP	VP NP
PP->IN NP	NP IN
VP->VBG PP	PP VBG
VP->VBD VP	VP VBD
VP->VBD S	S VBD
VP->VB PP	PP VB
VP->MD VP	VP MD
VP->VBP S	S VBP
VP->VBZ S	S VBZ
VP->VB NP	NP VB
VP->TO VP	VP TO
VP->VBG NP	NP VBG
VP->VBD ADJP	ADJP VBD
VP->VBZ VP	VP VBZ
VP->VBD PP	PP VBD
VP->VBN PP	PP VBN
VP->VBD NP	NP VBD
VP->VBZ NP	NP VBZ
VP->VBN NP	NP VBN
VP->VBP VP	VP VBP

A noun phrase comprises a noun and any associated modifiers: Verb phrases are composed of the verbs of the sentence and any modifiers of the verbs. Noun phrases are directly appended to the Reordered sentence. First part of the Verb phrase contains the verb of the sentence. This verb is extracted from the Verb phrase and is added to the end of the Malayalam sentence. Word reordering is performed using these techniques. The above example is illustrated in figure Figure 2 and Figure 3.

English Text (SVO):

He bought a watch for his wife

Reordered sentence in English (SOV):

He a watch his wife for bought

Malayalam Translation (SOV)

അവൻ ഒരു വാച്ച് അവന്റെ ഭാര്യക്ക് വേണ്ടി വാങ്ങി

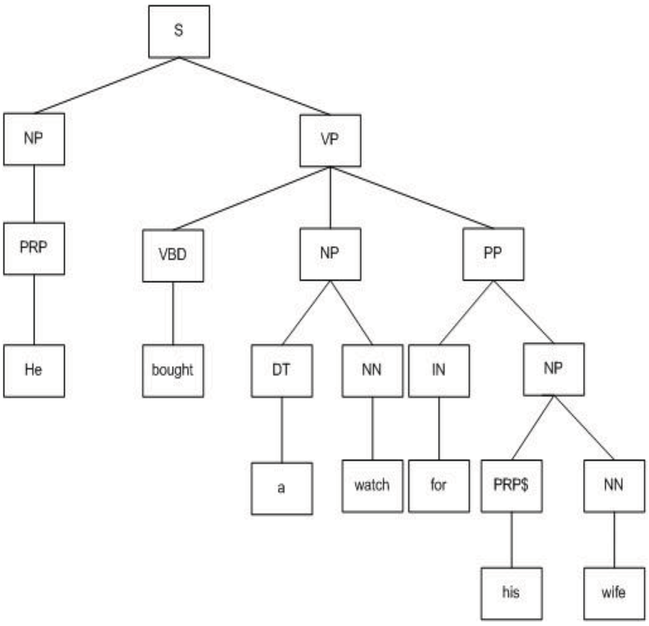

Figure 2. Syntax Tree for English text

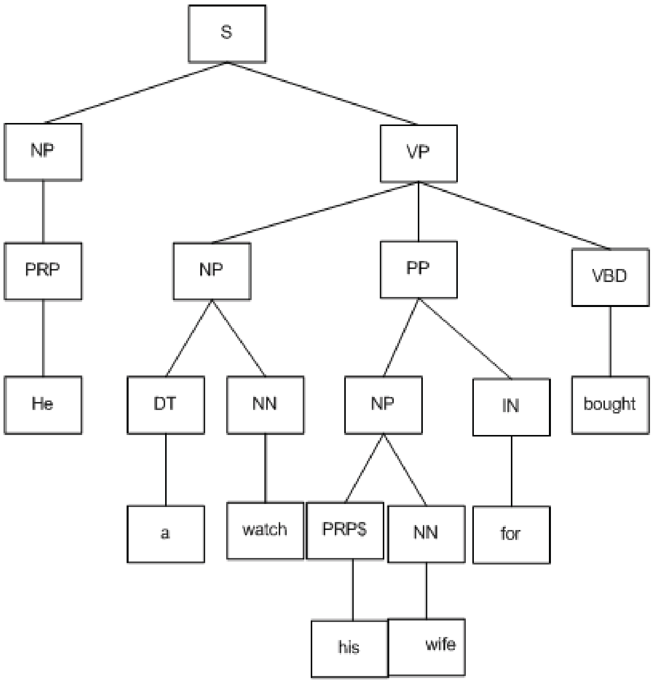

Figure 3. Syntax Tree for Malayalam text

C. Pattern Recognition

In this module various sentence patterns are identified based on the dependency information generated by the parser. Following Table III illustrates the patterns considered in this work.

TABLE III

ENGLISH SENTENCE PATTERNS

Pattern Structure	Example
Subject ,Verb	Fishes swim.
Subject, Verb ,Subject Complement	My father is a doctor.
Subject,verb,Direct Object	He knows me.
Subject,verb,indirect Object, Direct Object	She teaches us physics.
Subject, Verb, Direct Object, Proposition, Propositional Object	Ram bought a watch for his wife.
Subject,verb, Noun/Pronoun ,Adjective	I found the tin empty.
Subject,verb,Proposition ,Propositional Object	He believes in ghosts.
Subject, verb, to infinitive	We expected to win the match.
Subject,verb,Noun/Pronoun, to infinitive	His friends encouraged me to compete in the race.
Subject ,Verb, Gerund	They began playing.

General rules or tag sequences are identified based on the POS Tag information of the reordered sentence. According to each tag sequence certain case information is added to the POS tag.

Example:

SOV: He a watch his wife for bought

POS Tag PRP DT NN PRP$ NN IN VBD

Tag sequence is IN followed by NN. The NN is replaced with NNDC.

NN IN → NNDC INFR

NNDC is the NN in the dative case.

INFR implies, the preposition ‘for’ is used. in this text.

D. Translation

Reordered sentences are translated using a bilingual English Malayalam dictionary. The translation unit also uses a morphology generator which contains the transformation rules for Nouns, Pronouns and Verbs. Transformation rules are taken from [6].

1. Bilingual dictionary

A bilingual dictionary is used to translate words from English to Malayalam. SBMT system performs word by word translation based on the dictionary [14]. Only base form of the root word is stored in this. Malayalam words are stored using the Unicode format. The size and quality of dictionary limits the scope and coverage of a system.

2. Morphology Generator

Morphology generator [8] analyses the internal structure of the translated text. Malayalam is an agglutinative language in which a word is formed by adding suffixes to the root. Nouns are linguistic categories, which takes cases and PNG (Person, Number and Gender) information. Nouns change their forms according to Case (Vibhakthi), verbs change their forms according to the TAM (tense, Aspect, Mood). Person can be first person, second person or third person. Number is a part in an utterance, which indicates whether on object we talk about, is one, or more than one. Gender in language is the same as the universally known divisions of masculine, feminine and Neuter.

Vibhakti is the ending suffixes that are added to Nouns indicate their relationship with the central verb. Different cases of Malayalam language are Nominative, accusative, sociative, Dative, genitive and Instrumental. New tag generated in the pattern recognition phase is used to identify the different case information. Transformation rules given in [6] are used for case endings. Base form of the verb is stored in the bilingual dictionary. Stemming operation is performed on the main verb.. Different tense forms are generated using the transformation rules for verbs in [6].

Example:

Input Text: He a watch his wife for bought

New Tag sequence: PRP DT NN PRP$ NNDC INFR VBD

In this example case information for word wife is NNDC. i.e. NN in the dative case. Word ending character for the Malayalam translation of wife is a consonant. According to the translation rule the case marker is added to the Malayalam translation of wife. Past form of verb is present in the text.

Translated Text after adding case markers:

അവൻ ഒരു വാച്ച് അവന്റെ ഭാര്യക്ക് വേണ്ടി വാങ്ങി

IV. PERFORMANCE EVALUATION

Various methods for the evaluation for machine translation have been employed. Performance of the SBMT system was measured using Word Error Rate as well as the F-measure.

Word Error Rate (WER) is a common metric of the performance of a machine translation system. This problem is solved by first aligning the recognized word sequence with the reference word sequence using dynamic string alignment. Word error rate can then be computed as:

$$WER = (S + D + I) / N \qquad (1)$$

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of words in the reference

F-measure is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. Precision and recall are two widely used metrics for evaluating the correctness of a pattern recognition algorithm. Precision p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall where an F_1 score reaches its best value at 1 and worst score at 0.The overall performance of the system is summarized in the Table III.

$$F = 2\times((precision\times recall)/(precision + recall)) \tag{2}$$

TABLE III

PERFORMANCE OF THE SYSTEM

Sentence patterns	WER	F-measure
Unseen Sentences (including prepositional objects)	0.429	.0.57
Unseen Sentences	0.333	0.66

IV. CONCLUSION AND FUTURE WORK

In this paper a new approach for English to Malayalam translation is presented .This SBMT system identifies a set of reading rules that can be used for popular sentence patterns in English. This is a Syntactic Based Machine Translation in which different syntactic information about the source language is used for translation. As compared to other approaches, it is easy to identify the common rules for a group of sentences. These rules can be easily identified by analyzing the Syntax tree for the Malayalam text. This approach may improve the performance and effectiveness of translation.

In the future research more generalized rules can be obtained by incorporating more source language patterns. Translation efficiency may be improved by adding Aspect and Mood information of a verb. Work can be extended to handle complex and compound sentences.

REFERENCES

[1] Sinha, R. “ANGLABHARTI - A Multilingual Machine Aided Translation from English to Indian Languages”. IEEE Intl Conference on Systems, Man and Cybernetics (1995)

[2] Probst, K., Brown, R., Carbonell, J., Lavie, A., Levin L,and Peterson E. “Design and Implementation of Controlled Elicitation for Machine Translation ofLow-Density Languages”. In Workshop MT2010 at MachineTranslation Summit VIII.

[3] R.M.K. Sinha, A. Jain “AnglaHindi:An English to Hindi Machine-Aided Translation System”. Indian Institute of Technology, Kanpur, India, 2003.

[4] D. Chiang. “A Hierarchical Phrase-Based Model for statistical Machine Translation”. In Proc. of ACL, Ann, Arbor M ,2005.

[5] R. Zens and H. Ney “Discriminative Reordering Models for Statistical Machine Translation”. In Proc. of HLT-NAACL Workshop on SMT, New York, NY 2006.

[6] Sumam Mary Idicula and Peter S David. “A Morphological processor for Malayalam Language,South Asia Research”.SAGE Publications,2007.

[7] Kay Rottmann, Kay Rottmann and Stephan Vogel. “Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model”,2007.

[8] Rajeev R R, Elizabeth Sherly. “ A suffix Stripping based Morph Analyser for Malayalam Language”. Proceedings of 20th Kerala Science Congress , p 482-484, 28-31, (2008).

[9] Ananthakrishnan R, Hegde J, Bhattacharyya P, Shah R, Sasikumar M. “Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation”. In the Proceedings of International Joint Conference on NLP (IJCNLP08), Hyderabad, India (2008).

[10] Jakob Elming, Nizar Habash. “Syntactic Reordering for English-Arabic Phrase-Based Machine Translation”. Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 69–77.

[11] Remya Rajan, Remya Sivan, Remya Ravindran, K.P Soman, ”Rule Based Machine Translation from English to Malayalam”. IEEE International Conference on Advances in Computing,Control and Telecommunication Technologies 2009.

[12] Mary Priya Sebastian, Sheena Kurian K. and G. Santhosh Kumar. “A Framework of Statistical Machine Translator from English to Malayalam”.Proceedings of Fourth International Conference on Information Processing, Bangalore, India (2010) .

[13] Mohammad M.Abu Shquier,Mohammed M. Al Nabhan Tengku Mohammed Sembok. “Adopting new Rules in Rule-Based Machine translation”. 12th Intrnational Conference on Modelling and Simulatio 2010.

[14] Jisha P.Jayan ,Rajeev R R, Dr. S Rajendran. “Morphological Analyser and Morphological Generator for Malayalam-Tamil Machine Translation” International Journal of Computer Applications (0975 – 8887) Volume 13– No.8, January 2011.

[15] Sweden.R. Brown. “Example-based machine translation in the pangloss system”. Proceedings of the16th International Conference on Computational Linguistics (COLING-96), pages 169–174.