IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay

MONASH University

Project Title:	**Improvement in current MT system by incorporating hierarchical factor based MT**

Project Number	**IMURA0371**	

Monash Supervisor(s)	Main Supervisor: **Reza Haffari**	*Full names and titles*

Monash Primary Contact:	reza@monash.edu	*Email, phone*

Monash Head of Department:	Kim Marriott kim.marriott@monash.edu	*Full name, email*

Monash Department:	Clayton School of IT	*Full name*

Monash ADRT:	Kai Ming Ting KaiMing.Ting@monash.edu	*Full name, email*

IITB Supervisor(s)	Main Supervisor: **Pushpak Bhattacharya**	*Full names and titles*

IITB Primary Contact:	pb@cse.iitb.ac.in	*Email, phone*

IITB Head of Department:	A Sanyal as@cse.iitb.ac.in	*Name, Email,*
IITB Department:	Computer Science and Engineering	*Full name*

Research Academy Themes:

Highlight which of the Academy's Theme(s) this project will address?

Advanced computational engineering, simulation and manufacture

The research problem

Currently tree based MT and factor based MT rule the MT scenario. Both work on top of phrase based MT. While the former is more syntactically motivated the latter is linguistically motivated system. We will consider them one by one.

Tree based MT claims to open a linguistic understanding of machine translation. Basically a tree represents linguistic feature of the language. Based on which side is represented linguistically various models came into existence.

Some of the approaches are hierarchical based MT, Tree-to-string MT, String-to-Tree MT, Syntax augmented MT and syntax directed MT. These approaches have coexisted primarily due to researchers reluctance on meeting a common front to define Tree based MT. Syntax augmented MT and string to tree based MT are linguistic in the target side. Syntax directed MT, tree to string are linguistic in the source side. On the other hand hierarchical based MT is linguistic in neither side.

On the other hand factor based MT works on top of phrase based MT by incorporating linguistically motivated factors like stem, part-of-speech and other morphological features, it removes ambiguity and gives better quality translation. But still the disadvantages associated with phrase based MT persists. It cannot handle translation of discontiguous phrases. So the obvious step would be to combine the strength of hierarchical structures and factors and propose a new MT technique called hierarchical factor based MT.

Project aims

Compromise and optimise
IT is interesting to find out how to amalgamate the two approaches to produce a better translations. Also a lot of optimisation will be required to reduce time complexity. Apart from optimisation, one also has to take care that pruning does not reduce quality of translation. Factor based hierarchical model must be developed keeping optimisation and quality assurance into consideration. The best compromise between the two is what this proposal aims at achieving.

Expected outcomes

Milestones
1. 8-9 months Making a theoretical foundation of the factor based hierarchical model .
2. 8-9 months Programming the model into a software package.
3. 5-6 months Making an end to end model which takes a parallel corpus and generates a BLEU score on a testset.
4. 5-6 months Evaluating on variety of corpus keeping hierarchical model of Joshua MT system as base case.

How will the project address the Goals of the above Themes?

While tree based MT takes care of reordering due to language divergence, factor based MT takes care of linguistic sparsity in data by delving into morphological aspect of language. Tree based MT makes use of transfer rules to convert tree in source language to target language. This solves some of the reordering problems associated with phrase based MT. But the ambiguity in language cannot be solved by tree based MT alone. Factor based MT resolves the ambiguity by analysing source language at the morphological level in the Vauquois triangle. Using a knowledge

base it transfers the source language word into target language word keeping the morphological aspects of the source language words like POS, number, gender, tense intact.

Approach
In factor based MT factoring was done at word level. Now that factors have to be computed instead of just phrases and this has to be done prior to decoding. So the decoding algorithm of tree based MT basically remains same, only computations increase due to factors. So each possible translation option is expanded with factors. The option that has the highest score given by a log linear model over all the features is considered to be the best translation. But one problem could arise due to large number of possible translation options. Each hypothesis will have multiple mapping. This must be handled by approximate algorithms like pruning, that maintain a list of top N hypotheses at each expansion.

In doing so, quality translations must be retained. So a reranking algorithm will be required such that required results are in top. In case a slightly ungrammatical sentence is obtained, small grammatical errors could be removed using a grammatical error removal module.So pruning a lot will reduce space and pruning less will increase time complexity. The best result is one that optimised translation result.

Case studies

Here i have compared the translations of hierarchical model of Joshua MT system and factor based model of Moses MT system. These translations shows the need for a new model because both the translations lack in fluency and adequacy.

1.*Hindi:-* *आज*

और *पर* *अब*

।

English:- It is believed that before the formation of modern Delhi, Delhi was ruined seven times and established in different areas, some of the remains can be seen today also.
Hierarchical model (Joshua):- today it is believed that modern Delhi seven times before it became Delhi and sprang at various places which remains can still be seen.
Factor based model (Moses):- it is believed that the modern Delhi today before Delhi seven times and various places, some of whom unexciting remains can be seen even now.

2.Hindi:-

।

English:- Tulsidas is also considered to be an Avatar(incarnation) of Sage Valmiki who wrote the original epic Ramayana.
Hierarchical model (Joshua):-Tulsidas is regarded as the incarnation of the Maharshi Valmiki like the original poetry the author of Ramayana .
Factor based model (Moses):-Tulsidas, the Maharshi Valmiki is considered as an incarnation of the original,etc . , which was the author of the Ramayana poetry.

Capabilities and Degrees Required

Minimum requirement is Master's Degree in CSE and is working in the field of hierarchical statistical MT and factor based SMT.