IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project Title:	**Statistical Machine Translation for resource poor languages**	
Project Number	**IMURA0412**	
Monash Main Supervisor (Name, Email Id, Phone)	Prof. Reza Haffari Gholamreza.Haffari@monash.edu	*Full name, Email*
Monash Co-supervisor(s) (Name, Email Id, Phone)		
Monash Head of Dept. (Name,Email)	Prof. Graham Farr	*Full name, email*
Monash Department:	Clayton School of IT	
Monash ADRT (Name,Email)	Prof. Kai Ming Ting	*Full name, email*
IITB Main Supervisor (Name, Email Id, Phone)	Prof. Pushpak Bhattacharyya pb@cse.iitb.ac.in	*Full name, Email*
IITB Co-supervisor(s) (Name, Email Id, Phone)		
IITB Head of Dept (Name, Email, Phone)	Prof. S. Sudarshan head@cse.iitb.ac.in	*Full name, email*
IITB Department:	Computer Science and Engineering	

Research Academy Themes:

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

(Feel free to nominate more than one. For more information, see www.iitbmonash.org)

1. Advanced computational engineering, simulation and manufacture
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research

The research problem

Define the problem

Large amounts of parallel bilingual text is a necessity for training and building high quality statistical machine translation (SMT) models. However, big bilingual parallel text is not available for many world language pairs whereas comparable bitext may be available. In this project, we aim to overcome the scarcity of parallel bitext for language-pairs which belong to the same language family. In particular, we consider Hindi-Farsi language pair and the problem of building a statistical machine translation system between these two languages for which we do not have access to bilingual parallel data.

Project aims

The main aim is to build a high quality translation system between two languages for which we do not have enough parallel sentences. To achieve this main goal, we consider the following sub-goals:
(1) automatic mining of bilingual lexicon from comparable corpora
(2) improvement of reordering model and the morphological treatment leveraging the followings

- *The fact that the two languages, in our case Hindi-Farsi, belong o the same family*
- *The existence of a bridging language. For Hindi-Farsi, we have access to some parallel bilingual data for English-Hindi and English-Farsi. Therefore, it is potentially possible to use English to brigdge the gap between Hindi and Farsi.*

Expected outcomes

- *the development of statistical frameworks*
- *the development of algorithms*
- *C++/Python code for implementing the approach*

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

To achieve the project goals, our approach consists of the followings:

- We will devise statistical models to leverage the similarity of the languages in syntax for a better reordering when translating from one language to the other
- We make use of “comparable” bitext for creating a better bilingual lexicon (aka dictionary) to inject to the translation system
- We leverage the similarity between the two languages in morphology, and capture by appropriate statistical model.
- We leverage the existence of a bridging language.

While English has excellent language tools and resources and Hindi too has reasonable amount (through the long standing NLP effort in India), Persian is not far behind. Persian is a heritage language with enormous implication for tourism, commerce and other domains. Following form a sound platform to leverage for our project:

Parallel resources

- http://translate4iran.wikispaces.com/Khyaban+4.2
- http://ece.ut.ac.ir/NLP/ resources.htm (1,50,000 sentence pairs extracted from subtitles. Roughly 4M tokens on each side. Available freely by requesting through e-mail)
- http://catalog.elra.info/product_info.php?products_id=1111 (about 100,000 sentences provided by

ELRA). The price varies from 500 to 3000 euros depending on the intent (research/commercial etc.)

http://www.centcom.mil/

Monolingual resources

- http://ece.ut.ac.ir/DBRG/hamshahri/download.html (Monolingual corpus from news domain. Claims to be the largest monolingual corpus. Free download)
- http://www.lingoistica.com/news/list (can crawl this to get some amount of Monolingual corpus from news domain.)

Persian Treebank
http://stp.lingfil.uu.se/~mojgan/UPDT.html (6000 sentences)

Tools
This page has some tools (tokenizer, normalizer, part of speech tagger, parser)
http://stp.lingfil.uu.se/~mojgan/

Capabilities and Degrees Required

The student should be fluent in advanced statistics, algorithms and data structure, and coding. It will be an advantage if he/she is good at spotting language phenomena and likes linguistics.

Potential Collaborators

Primary collaborators:
Prof. Reza Haffari, Monash & Prof. Pushpak Bhattacharyya IITB

Secondary collaborators:
Graduated PhDs from IIT Bombay NLP group, working in IBM, Samsung, Microsoft etc. are potential collaborators.