Creation of Machine Translation tools and resources for English to Dravidian Languages: Pilot Study

Our main objective is to develop Machine Translation (MT) system and needed linguistic resources for English-Dravidian languages (Tamil, Malayalam, Telugu and Kannada), that would facilitate the creation of rich educational contents in Indian languages. Our research effort is to make all the tools and translation system to be based on Machine Learning methodologies so that computer graduates and other such non-linguists are able to immediately participate in the national mission on literacy by contributing additional tools for language translation.

We have divided the project into 3 modules:
Module I : Aims at developing Machine translation tools and linguistic resources for English to Dravidian languages. We will also work in creating Machine Translation tools for Malayalam to Tamil and vice versa. This later can be extended to other language pairs.

Module II : We aim at developing teaching material corresponding to the tools we are developing (most tools are based on machine learning) so that it can be delivered as part of undergraduate computer science and engineering curriculum on data mining/machine learning. This will ensure a critical amount of man power required for sustaining translation effort needed for national mission on education.
Module II also aims at training 500 faculties selected from across the country on machine translation methodologies using machine learning techniques.

Module III : Aims at developing a Dravidian Wordnet required for translation. This module requires collaboration between different universities. This will also link to the Hindi wordnet developed at IIT Bombay (www.cfilt.iitb.ac.in) which is being widely used for NLP involving Indian languages.

SUMMARY

The main objective of DWN is to develop an extensive and high quality multilingual database with Wordnet for Dravidian languages in a cost-effective manner, and also to link it with the existing prominent Wordnet of languages like Hindi, English and other languages. The project will also develop a language independent set of semantic concepts linking the language networks together. The resources will be field tested for adequacy in an information retrieval application. The ultimate objectives are to move toward standardization of semantic classification of information for all Dravidian languages and to provide resources for development of applications, which can operate in a selected language or over a range of languages.

Major aims of the project are:

1. The major aim of the project is to prepare Wordnet for the following languages:

   - Tamil Wordnet
   - Malayalam Wordnet
   - Telugu Wordnet
   - Kannada Wordnet
   - Linking of Dravidian Wordnet with the Wordnet of Hindi and other languages in North India and also with English Wordnet

2. These individual Wordnet will be merged into Dravidian Wordnet and also linked with Wordnet of English and Hindi.

3. The notion of a synset and the main semantic relations will be taken over in Dravidian Wordnet. However, some specific changes will be made to the design of the database, which are mainly motivated by the following objectives:

   1. to create a multilingual database;
   2. to maintain language-specific relations in the Wordnet
   3. to achieve maximal compatibility across the different resources; to build the Wordnet relatively independently (re)-using existing resources

Scope of work

Two types of tools are proposed in this project.

1. Machine Translation tools
2. Computer Assisted Translation tools

These two technologies are the consequence of different approaches. They do not produce the same results, and are used in distinct contexts. MT aims at assembling all the information necessary for translation in one program so that a text can be translated without human intervention. It exploits the computer's capacity to calculate in order to analyze the structure of a statement or sentence in the source language, break it down into easily translatable elements and then create a statement with the same structure in the target language. It uses huge plurilingual dictionaries, as well as corpora of texts that have already been translated. As mentioned, in the 1980s MT held great promises, but it has been steadily losing ground to computer-assisted translation because the latter responds more realistically to actual needs.

BACKGROUND AND TECHNOLOGY STATUS

Comparison with existing MT systems in India

There are mainly three MT systems (TAG, Analgen and SMT) that are being developed by member institutions of the EILMT consortia. The TAG(CDAC-Pune) and Anlagen(IIIT- Hyderabad) works adequately for English to Hindi translation. However the peculiar nature of Dravidian languages demands further fine tuning and research.

So we made an attempt to develop a new system. A prototype system is already developed which gives encouraging results. In parallel, we are also working on Machine learning based translation system, such as source tree to target string mapping and machine learning model that maps parse trees in the source language to parse trees in the target language (model is to be learned from parallel corpus).

We also propose to combine the merits of rule based and Machine learning based system.

Resources required for translation between Indian languages

1. Morph analyzer and tagger
2. Parser
3. Transliteration
4. Morph synthesizer
5. Wordnet

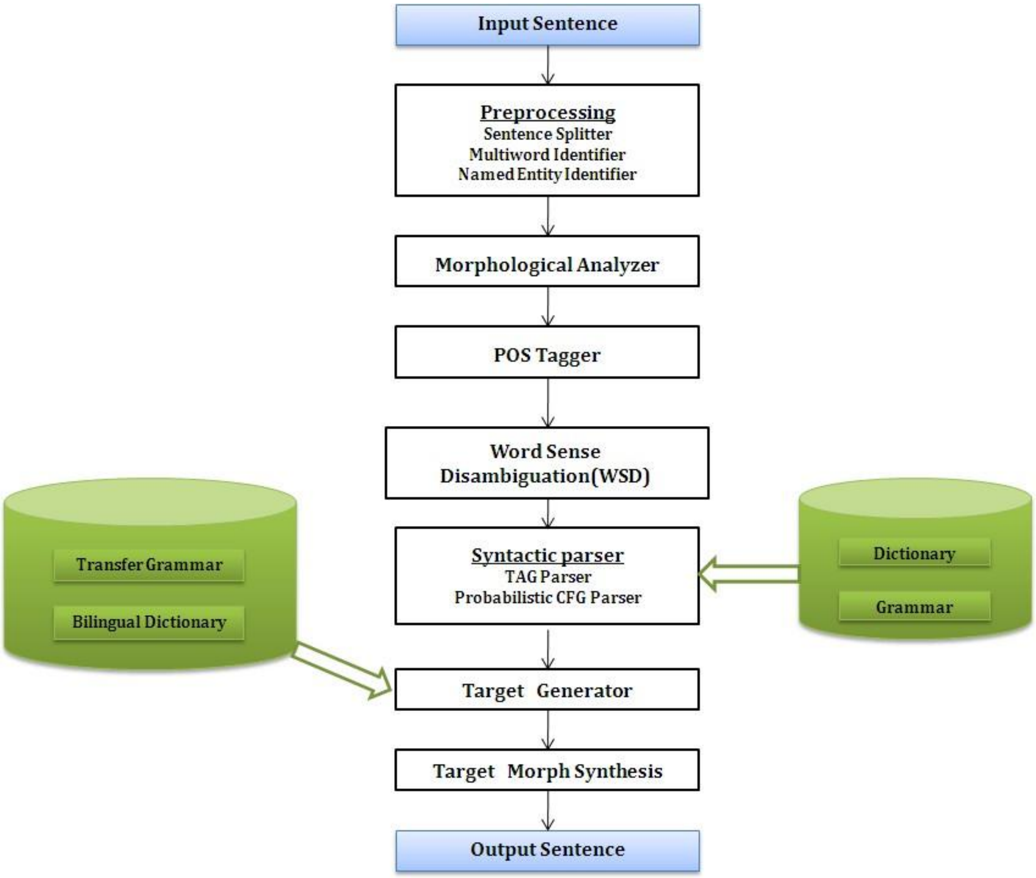

Figure 1: Architecture of the Rule based MT System

Our rule based architecture belongs to syntax transfer approach as shown in Figure 1. Not all the modules in the architecture are rule based, tools such as Part of Speech (POS) tagger and parser are statistical ones. Our approach effectively focuses on developing target resources rather than focusing on English. i.e. we will be using state of the art resources available for English in our system, this will reduce development time. On the syntactic parser side, we have two options. One is, parsing the input sentences using Tree Adjoining Grammar (TAG) parser and another is parsing the sentences using probabilistic Context Free Grammar (CFG) parser. For English- Indian language MT, parsers available in the public domain will be used. For Indian-Indian language MT, we will be developing parser and grammar resources on our own which are based on machine learning.

EXPECTED OUTCOMES:

One of the main objectives of the mission is to make the educational contents available in Indian Languages. English being Lingua-franca of science and technology, most of educational materials are being created in world over is in English. The aim of this project is to develop tools and resources that assist in translating such materials.

Translation on its own is a nascent industry in India. Tools and resources developed in this project will give a boost to this industry giving jobs to thousands in the country.

DELIVERABLES YEAR WISE AND ITS POSSIBLE CONTRIBUTION TO MAJOR OBJECTIVES OF MISSION.

Module I Phase 1 (12 months from the start), the following will be delivered

1. POS Taggers and tagged data for all Dravidian Languages
2. Morph Analyzer for all Dravidian Languages version 1
3. Transliteration Engine for all Dravidian Languages.
4. Translation Engine for English-Tamil – (Hybrid of Rule and Statistical)
5. Translation Memory system for manual translation
6. Parallel corpora (5000 sentence pairs) for all Dravidian languages

Module II: Phase 1 (12 months from the start), the following will be delivered

1. Text resources on Machine Learning based Computational Linguistics of Indian Languages.
2. Video Resources: Video Lectures of the key concepts in the textbook.
3. Lab Assignments with data sets.

Module III: Phase 1 (12 months from the start), the following will be delivered

1. First version of Wordnet for each language except Malayalam

TIME SCHEDULE (YEAR WISE)

Tasks	Start Month	Duration	End Month
Hiring people, Purchasing hardware	0	1	1
Corpus Generation for Linguistic tools based on Machine Learning	1	4	5
Training and Tuning the Learning tools	5	7	12
Creation of Parallel Corpora for all Languages	1	11	12
Translation Engine for English-Tamil	1	9	10
Testing of Translation engine	9	3	12
Creation of Text Resources	1	11	12
Creation of Video Resources	7	5	12
Wordnet Creation	1	11	12

PROPOSED BUDGET

(a) Recurring budget (not more than 30%) of the proposal along with item-wise breakup (Man power, Contingency, Consumable, Travel, Miscellaneous year wise breakup.

Total Recurring Budget

No:	Task	Amount in Lakhs
1	Manpower	50.7
2	Honorarium	6.4
3	Contingency	12.00
4	Travel	10.70

Total Amount: Rs. 79.8 Lakhs

(The project is highly manpower intensive and hence recurring budget is more than non-recurring budget)

(b) Detailed breakup of non-recurring items (with the equipment to be procured along with cost)

Total Non-Recurring Budget

No	Task	Amount in Lakhs
1	Equipment	20.2

Total Amount: Rs. 20.2 Lakhs

Total Proposed Budget (recurring and non-recurring): Rs 100 Lakhs

REFRENCES

i. Rajat Mohanty and Pushpak Bhattacharyya, *Lexical Resources for Semantic Extraction*, Lexical Resources Engineering Conference (LREC08), Marrakech, Morocco, May 26-June 1, 2008.

ii. Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra and Aditya Sharma, *Synset Based Multilingual Dictionary: Insights, Applications and Challenges*, Global Wordnet Conference (GWC08), Szeged, Hungary, January 22-25, 2008.

iii. J.Ramanand, Akshay Ukey, Brahm Kiran Singh and Pushpak Bhattacharyya, *Mapping and Structural Analysis of Multilingual Wordnets*, IEEE Data Engineering Bulletin, 30(1), March 2007

iv. Manish Sinha, Mahesh Reddy and Pushpak Bhattacharyya, *An Approach towards Construction and Application of Multilingual Indo-Wordnet*, 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006.

v. Rajat Mohanty, Anupama Dutta and Pushpak Bhattacharyya, *Semantically Relatable Sets: Building Blocks for Repesenting Semantics*, 10th Machine Translation Summit (MT Summit 05), Phuket, September, 2005.

vi. G. Ramakrishnan, Krishna Prasad Chitrapura, Raghu Krishnapuram and Pushpak Bhattacharyya, *A Model for Handling Approximate, Noisy or Incomplete Labeling in Text Classification* , International Conference on Machine Learning (ICML 05), Bonn, August, 2005.