Statistical Relational Learning for Machine Translation

IIT Bombay

What is Statistical Relational Learning?

SRL is a sub discipline of Machine Learning which borrows heavily from Probability, Statistics and First Order Logic.
It is used in domains which are uncertain and have a complex relational structure.
Probabilistic Graphical methods like Bayesian Networks, Markov Networks are used to model the uncertainty while First Order Logic is used to describe the relational properties by universal quantification.
This field doesn’t focus only on learning but also on reasoning and knowledge representation.

Project Description

Over the last few years, we have witnessed statistical machine translation (SMT) approaches becoming popular.

They have provided the technology for building broad coverage machine translation systems for multiple languages with fast development and deployment time.

The methods that underlie these data driven approaches are primarily geared toward ensuring adequacy of translations.

Good adequacy is generally assured by SMT systems but fluency is not. However, these aren’t independent to the human evaluator.

This involves ensuring correctness of the constituent structure of the translation, case marking, morphology and agreement among other issues.

Project Description (Continued)

SMT research is focused more towards constituent structure while the others, which are important phenomena in Indian languages, are receiving less attention.
Constituent structure is also an important factor while translating from English to Indian Languages.
India, with all its diversity, needs content available in local languages.
We will be translating books from English to Indian languages with professional editors as human evaluators.

Interesting research directions

Usage of resources and methods for learning SRL Models on the MT output and capture of error correction patterns based on input from editors. These editorial inputs could be at various levels of granularity: from routine correction of certain sentences to non-trivial corrections at the discourse level.

Effective utilization of monolingual resources like corpora (in the SRL models) which are currently underutilized.

Factoring in mistakes made by NLP tools as well as the SMT systems in the SRL models.

Broad Objectives

To speed up translation of books.
To develop superior set of SRL models and rules, based on the learnings from the human professional editorial inputs, across books of the same genre.
Over time to create a broader set of rules which could be applicable across different genres of books.
To apply, if possible with the concomitant limitations, the above learnings across languages.
Making Indian language content available on the Internet to make Internet based searches in Indian languages as satisfactory as in English.

Approach

To avoid delay, we intend to acquire user rights to an “off the shelf” machine translation software, preferably with its programming interface.
Initially, we will translate English books to only Hindi, Gujarati and Tamil. We might expand to other languages depending on success.
We intend to avoid literary books and focus more on developing a corpus in the field of Law, Economics and Sciences.