Detecting Complex Predicates in Hindi across Parallel corpora

CS671: Introduction to Natural Language Processing
Under Guidance of Prof. Amitabh Mukerjee

Project Proposal

Group X : Amit Kumar (10074), Ankit Modi (10104)

Introduction

Given a parallel English-Hindi parallel corpora, we have to detect all kinds of complex predicates (CPs).
This approach uses the fact that a CP is a multi-word expression with its meaning being distinct from the light verb (LV). A *Light Verb* is defined as a verb that has little semantic content of its own and it therefore forms a predicate with some additional expression, which is usually a noun.

Motivation

Hindi has an abundance of CPs. CPs improve the expressiveness of a language and their detection is a tough task. Their detection would provide an important resource for various tasks such as constructing Wordnets, linguistic analysis of CPs, information retrieval, summarization etc.

Related Work

This problem was solved using word alignment and POS tagging of parallel sentences [2]. In this approach POS tagging of English sentence is projected on the parallel hindi sentence. Then by using word alignment of both sentences CPs are identified.
CPs had been mined using computational methods and then, were categorized using statistical analysis (*Sriram and Joshi, 2005*).
Chakrabarti et al (2008) present a method for automatic extraction of CPs only from a corpus based on linguistic features.
Derivation of complex predicates has also been dealt with linguistically and computationally [3]. Authors have constructed empirical tests to decide if a combination of two words, the second of which is a verb, is a complex predicate or not. Such tests will provide a principled way of deciding the status of complex predicates in Indian language wordnets.

Implementation

Our approach is mainly based on to [1] .

Input for our project will be parallel sentences in English and Hindi languages.
Main steps of our approach :

1. Search for the light verb (LV) present in the hindi sentence.
2. If the English meaning of this LV is present in the parallel English sentence then exit.
   else scan the hindi sentence and get the word(W) in the left of LV.
3. IIf W is an exit word then EXIT and start again from step 1.
   else If the English meaning of W is present in the English sentence then W+LV is our Complex Predicate.
   else update W to the word on it's left in hindi sentence.

Resources

- Corpus: We are looking for a Hindi-English parallel corpora
- List of Hindi Light Verbs: *Reverse Complex Predicates in HIndi* by Shakthi Poornima, Department of Linguistics, SUNY University at Buffalo
- Morphological forms of English verbs: http://www.englishpage.com/irregularverbs/irregularverbs.html

References

- [1] *Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus*, R. Mahesh K. Sinha, IIT Kanpur
- [2] *Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora,* Amitabha Mukerjee, Ankit Soni and Achla M Raina, IIT Kanpur
- [3] Complex Predicates in Indian Languages and wordnets. Pushpak Bhattacharyya, Debasri Chkrabarti and Vaijayanthi M. Sarma. Language Resources and Evaluation 40(3-4):331-355
- Wikepedia:
  1. http://en.wikipedia.org/wiki/Light_verb
  2. http://en.wikipedia.org/wiki/Compound_verb