IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project title

Disambiguation and Multilinguality Enabled Sentiment Analysis

Project number: IMURA0097

Monash University supervisors: Professor Ingrid Zukerman
Monash University contact: Professor Ingrid Zukerman; Email: ingrid@csse.monash.edu.au

IITB supervisors: Professor Pushpak Bhattacharyya
IITB contact: Professor Pushpak Bhattacharyya; Email: pb@cse.iitb.ac.in

The research problem

We propose the following tasks:

(a) Enhance sentiment analysis and opinion mining (SA & OPM) systems with sense and attachment disambiguation capabilities.
(b) Put in place tools and resources for SA & OM when sentiments and opinions are expressed in multiple languages with inevitable code mixing, i.e., mixing of languages.

Project aims

A crucial first step in SA & OM is good quality feature engineering on the documents expressing sentiments and opinions. The word tokens in the documents serve as first cut features. However, marked improvement in performance results, when this feature set is pruned retaining only the *charged words*, i.e., words with high polarity value. Such words are mainly adjectives, but other parts of speech words too qualify as deciding features (e.g., the verb *hate*). A very useful resource in this context is the Sentiwordnet which is the wordnet with polarity values associated with the synsets.

However, the use of this recourse needs word sense disambiguation. Our prior experience shows that in absence of disambiguation, the highly charged words loose their edge as deciding factors, for, one has to settle for some kind of average over the synsets in which such words participate.

WSD is known to be a hard problem; however, our prior experience again shows that cognizance and use of domain knowledge can achieve high WSD accuracy.

Attachment ambiguity refers to ambiguities resulting from PP attachment and scopes of adjectives and quantifiers. Wrong attachment and scope interfere with the detection of critical features.

The above points should suffice to bring to the fore the importance of our line (a) of investigation.

Coming to line-of-investigation (b), Multilinguality is a way of web-life today. Our prior work with corpora expressing sentiments in the *bank domain* in India showed that people freely mix code, i.e., mix languages in trying to forcefully make their points. Strong words from Hindi are borrowed in English blogs for effective expression. So is the case of borrowing English words when opinions are expressed in Indian languages. Handling code mixing is thus an important dimension of the problem.

Finally, there is a cross lingual IE and IR aspect to the whole task. Queries on what people think of an organization or individual should retrieve not only documents in the language of the query, but those in other

languages too. The information- needless to say- should be presented in the language of the query. It will be pertinent to mention in this context that at IITB, high level competence exists in multilingual computing.

Expected outcomes

In the light of the above discussion, the expected outcomes are:

1. SA & OM sensitive WSD solutions
2. Multilingual SA & OM tools and resources
3. Cross lingual SA & OM

Which of the above Theme does this project address?

Advanced Computational Engineering, Simulation and Manufacture.

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

Multilingual sentiment analysis is a relatively unexplored and challenging area which needs new, efficient and scalable solutions to address the needs of global internet community (Multinational Corporations are examples of them). The models and approaches learned during the course of project could well be used in other frontiers of natural language processing such as question- answering, analysing employee satisfaction etc.