Abstract

Human-machine interface technology has been investigated for several decades. In recent times, research activities in the areas of emotion in natural language texts and other media are gaining ground under the umbrella of subjectivity analysis and affect computing. The reason may be the explosive growth of the social media content on the Web in the past few years. We need a comprehensive theory of what a human emotion is, and then we need to understand how the emotion is expressed and transmitted within the natural language by incorporating syntactic, semantic, and pragmatic analysis of a text. Most of the current approaches address the problem of emotion analysis in English texts. This raises the demand of linguistic resources for languages other than English and cross-lingual study of emotions in natural language texts.

In order to obtain knowledge and information from emotional text, it is necessary to have reliable linguistic resources, such as tagged emotion corpora and emotion dictionaries. In case of English, two different types of corpus have been collected, a *SemEval 2007* news corpus and a blog corpus (Aman and Szpakowicz, 2007). The emotional expressions, sentential emotions and intensities are already annotated in the corpora but both the corpora do not contain any hints for emotion holder and topic. Thus, the annotation of these emotional components has been attempted for conducting different emotion analysis experiments in English. In case of Bengali, our annotation task addresses the issues of identifying emotional expressions in blog texts along with the sentential emotions, intensity, holders and topics.

In case of emotion lexicon in English, the *WordNet Affect* containing six types of emotion words in six separate lists is already available as open source. We developed the Bengali *WordNet Affect* from this English *WordNet Affect* lists. The automatic processes of updating, translation, sense disambiguation task and the inter translator agreement are among the important contributions in relation of the present thesis.

It is said that sentiment and/or emotion is typically a localized phenomenon that is more appropriately computed at the paragraph, sentence or entity level (Liu, 2009). Thus, the analysis of evaluative emotional expressions at word, phrase, sentence and document level granularities

has been attempted in the present task with various interesting insights in the experimental outcomes. The main focus of the present task is on phrase and document level emotion tagging as the word and sentence level emotion tagging has already been done in the Master's course. The document level emotion tagging has been carried out using two separate approaches, one is the machine learning based approach which assigns emotions to Bengali blog documents based on word to sentence and sentence to document level granularity and the other one is resource based approach that incorporates the emotion lexicon, Bengali *WordNet Affect* Lists (*BengWAL*) for identifying document level emotion tags.

An emotion holder is the person or organization that expresses emotion (Wiebe *et al.*, 2005). In case of identifying emotion holders, we developed two separate systems, a baseline system followed by a syntactic system. The baseline system identifies emotion holder in English based on the subject information from the typed dependency relations of the parsed emotional sentences. The parsing was done using the open source Stanford Dependency Parser. Similarly, for the morphologically rich languages (e.g., Bengali), we have used an open source Bengali shallow parser which produces different morphological information. We used the lexical pattern based phrase level similarity which contains different POS combinations, Name Entities (NEs) and noun phrases for identifying the emotion holders based on the morphological knowledge.

On the other hand, it is observed that the head of each chunk in the dependency-parsed output helps in constructing the syntactic argument structure with respect to the key emotional verb. Thus, two separate techniques have been adopted for identifying emotion holders from the syntactic argument structures of the emotional sentences, one is from the parsed result directly and another is from the corpus that has been POS tagged and chunked, separately. The hypothesis is that if the acquired syntactic argument structure of a sentence matches with any of the retrieved frame syntaxes of the English *VerbNet*, the holder roles (e.g., *Experiencer*, *Agent*, *Actor*, *Beneficiary* etc.) associated with the *VerbNet* frame syntaxes are assigned in the appropriate slots in the syntactic arguments. Similarly, in case of Bengali, each acquired syntactic argument structure with respect to its key verb is therefore mapped to all the possible frame syntaxes present for the equivalent English verbs of that key verb in the *VerbNet*. Additionally, we also address the roles of event actors and sentiment holders from the perspective of event sentiment relations within the *TimeML* framework.

In addition to emotion holder, emotion topic is a real world object, event or an abstract entity that is the primary subject of the emotion or opinion as intended by its holder (Stoyanov and Cardie, 2008a). We developed the baseline system for identifying emotion topics based on the object related dependency parsed relations. The phrase segments containing topic related *Thematic Roles* (e.g. *Topic*, *Theme*, *Event* etc.) are extracted from the verb based syntactical argument structures of the sentences. In addition to the baseline and syntactic systems, a supervised system is also built to identify multiple emotion topics along with their individual topic and target spans from each sentence. The Conditional Random Field (CRF), Support Vector Machines (SVM) and Fuzzy Classifier (FC) have been employed by considering various features (e.g., *the annotated emotional expressions along with direct and transitive dependencies, causal verbs, discourse markers, Emotion Holder, Named Entities* and four types of similarity measures like *Structural Similarity, Sentiment Similarity, Syntactic Similarity and Semantic Similarity*) and their combinations. The use of multi-engine voting technique on the output of the classifiers outperforms the baseline and syntactic systems.

The proper understanding of the emotion components and their associations is very important if we need to mine emotion properly from texts. It has to be mentioned that very few attempts have been carried out for identifying co-reference in case of opinion or sentiment analysis. Thus, we have shown how the building of fine-grained holder and topic knowledge based on rhetorical structure and segmentation of them using different types of lexical, syntactic and overlapping features substantially reduces the problem of the supervised framework. Thus, the identification of emotional co-reference is helpful in identifying user-topic relations because the co-reference information entails the presence of indirect affective clues that can be traced with the due help of holder and topic. The supervised system also shows the improvement over rule based baseline as the rule based system fails to capture the implicit textual clues whereas the supervised system captures the clues in terms of combined features. The evaluation of the co-reference using *Krippendorff's* alpha (2004) is helpful in diagnosing the importance of the three emotional components. The rule based post-processing techniques for reducing the error cases have shown substantial improvement in the performance of the system. In addition to identifying emotion co-reference, we have identified event and sentiment expressions at word level from the sentences of *TempEval-2010* corpus and evaluated their association in terms of lexical equivalence and co-reference.

The present thesis also highlights the tracking of emotions from blogs and events. Therefore, the identification of Ekman's six basic emotions from the bloggers' comments carried out at sentence and paragraph level granularities. The sense based affect scoring technique has been employed to identify the emotions. The *Self Affect Scores* (*SASs*) are used to measure the effect of the blogger's own previous emotions in identifying its present emotional state whereas the *Influential Affect Scores* (*IASs*) are considered for measuring the previous impact of all other bloggers' emotions in identifying a blogger's present emotional state. The change of a blogger's emotions has been tracked based on the emotions that are assigned to the nodes of the blogger's *Referential Informative Chain* (RIC). The evaluation techniques produce satisfactory performance in case of emotion tracking of the bloggers. On the other hand, the tracking of sentiment events based on temporal relations have been attempted in two forms, *sentiment twist* and *sentiment transition*. The *sentiment twist* identifies the change of sentiment between two consecutive events whereas *sentiment transition* identifies the change of sentiment in between more than two sentiment events.

Above all, the opinion or emotion labeled data for multiple languages is till-date a scarce resource. This is because the cost of assigning labels to all the data is expensive and time-consuming. Thus, we have prepared the emotion annotated data from unlabelled samples for multiple languages such as Bengali, Hindi, Telugu and Japanese. The *WordNet Affect*s and translated emotion tagged news corpora for these languages have also been prepared from the English *WordNet Affect* and *SemEval 2007* emotion corpus, respectively. It has been observed that a perfect sense to sense mapping among languages is impossible but the emotional senses do hold across languages. This implies that this information could be leveraged in an automatic fashion to provide additional clues for the affect labeling of unseen senses.