Subjectivity Detection in English and Bengali: A CRF-based Approach

Amitava Das and Sivaji Bandyopadhyay

Computer Science and Engineering Department

Jadavpur University, Kolkata-700032, India

Introduction

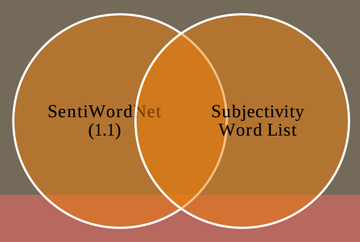
Extracting opinions from text is a hard semantic problem. Subjectivity Detection is studied as a text classification problem that classifies texts as either subjective or objective. This paper illustrates a Conditional Random Field (CRF) based Subjectivity Detection approach tested on English and Bengali multiple domain corpus to establish its effectiveness over multiple domain perspective. The final classifier has resulted precision values of 76.08% and 79.90% for English and 72.16% and 74.6% for Bengali for the news and blog domains respectively.

SentiWordNet (Bengali)

Two main lexical resources for English

- SentiWordNet
- Subjectivity Word List

✓ Merged new list
✓ Duplicate elements removed
✓ New list consist of 14,135 entries



	SentiWordNet		Subjectivity Lexicon	
Entries	Single	Multi	Single	Multi
	115424	79091	5866	990
Unambiguous Words	20789	30000	4745	963
Discarded Ambiguous Words	Threshold	Orientation Strength	Subjectivity Strength	POS
	86944	30000	2652	928

Statistics of Both English Sentiment Lexicon

Procedure

- Word Level Translation Model
- Samsad English-Bengali Dictionary has been used
- Filtration Techniques
  - Words with orientation low strength discarded
  - Words with undefined POS discarded
  - Filtration technique for Words lost Subjectivity after stemming
    - Stemming clusters
    - If Cluster center (root) has no sentiment orientation then the cluster has been discarded

  Ex: zeal, zealot, zealous, zealously

Theme Identification

Rule-Based Strategy

- Theme as a set of significant keywords in the document collection
- Significant Keywords identified using TF-IDF, Positional and Distribution factor
- Theme clusters, i.e., document set sharing theme words, identified
- Title words considered as high probable theme words
- Top ranked 5 significant words in each document as theme words

Document-level Theme Identification

Theme Clustering

Themes	Doc ID
ইরান,করিয়া,কর্মসূচি,কারণেই,আমেরিকা,পদক্ষেপ	Doc1, Doc78, Doc45, Doc135
তথ্যপ্রযুক্তি,ক্ষেত্র,শিল্প,পরিবহণ,সিটু,ধর্মঘটে	Doc22, Doc177, Doc37
দলিত,রাম,কাঁসি,মায়াবতী,রাজনৈতিক,নেতা	Doc32, Doc56, Doc79, Doc101, Doc83
পামুক,ওঁর,উপন্যাস,নোবেল,সাহিত্য,ওরহান	Doc12

Stemming Cluster

- Unsupervised Technique
- Simple suffix stripping for inflectional morphology
- Minimum Edit Distance for derivational morphology
- Calculate edit distance for insertion and deletion of 3 character (+3 to -3)
- Work on document level
- Manually generated suffix list

Stemming Clusters
সত্যজিৎ,সত্যজিৎকে,সত্যজিত,সত্যজিতের
ছবি,ছবির,ছবিটি,ছবিতে,ছবিতেই,ছবিটিতে,ছবিটির
রায়,রায়ের
ওপর,ওপরেও
করত,করতেন
নিয়ে,নিয়েছেন
দেন,দেননি
নির্মিত

Rule-Based Classifier

Procedure
✓The classifier first marks sentences bearing opinionated words. Every opinionated word is validated along with its POS tag in the developed SentiWordNet (Bengali).
✓Marks theme cluster specific phrases in each sentence.
✓In the absence of theme words, sentences are searched for the presence of at least one strong subjective word or more than one weak subjective word for its consideration as a subjective sentence.
✓The recall measure of the present classifier is greater than its precision value.

Supervised Classifier

Each document is represented as a feature vector for machine learning task. After a series of experiments the following feature set is found to be performing well as a subjectivity clue.

Features

Positional Factors	Percentage	
	MPQA	Bengali
First Paragraph	48.00%	56.80%
Last Two Sentences	64.00%	78.00%

Evaluation Result

Observations

- Subjectivity detection is trivial for review corpus and blog corpus rather than for news corpus
- Performance incremented by 2% only from rule-based and statistical (CRF)

Features	Overall Performance Incremented By	
	English	Bengali
Stemming Cluster	5.32%	4.05%
Part Of Speech	4.12%	3.62%
Chunk	3.98%	4.07%
Average Distribution	2.53%	1.88%
Sentiment Lexicon	6.07%	5.02%
Positional Aspect	3.06%	3.66%

Languages	Domain	Precision	Recall
English	MPQA	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	NEWS	72.16%	76.00%
	BLOG	74.6%	80.4%