Extracting Opinion Statements from Bengali Text Documents through Theme Detection

Amitava Das and Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
Kolkata-700032, India
Email:amitava.santu@gmail.com,
sivaji_cse_ju@yahoo.com

Data Acquisition from Web

- Bengali news corpus collected from web
- A focused web crawler to crawl the pages
- Corpus Cleaning
- Code conversion to get the data into standard unicode format
- Present work based on Reader's opinion section and Letters to the Editor Section in the news corpus

Corpus Statistics	
Total number of documents in the corpus	3,976
Total number of sentences in the corpus	87472
Avgerage number of sentences in a document	22
Total number of wordforms in the corpus	1145088
Avgerage number of wordforms in a document	288
Total number of distinct wordforms in the corpus	17176

Pre-processing the corpus

- Sentence Identification
- Tokenizer
- Part Of Speech (POS) Tagging
  - CRF based POS tagging engine
  - Focused on nouns, named entities, adjectives, adverbs and verbs

Training-Set	Test-Set	Accuracy
16397	4587	87.23%

- Chunking
  - CRF based chunking engine
  - Focused on NPs and VPs

Training-Set	Test-Set	Accuracy
16397	4587	79.51%

Theme Detection

- Theme Detection
  - Theme as a set of significant keywords in the document collection
  - Suffix striping with minimal string matching score for Theme Detection
  - 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 Retrieval

- Index File Creation
  - Theme word as Index keyword, Document Title, Document Id, Relevance
- Document Retrieval
  - Query type: Conjunction or Disjunction of Query words
  - Searching Index Files with Query (Theme) words
  - Documents with Query words in title get higher score
  - Identify all theme words in each retrieved document
  - Theme words for a particular query, i.e., the theme bag, is the union of theme words of the retrieved documents

Opinionated & Non-Opinionated Sentence

- Sentence identification using theme bag and
  - Linguistic rules based on POS and chunk information
  - Nouns and NEs among the theme words considered as aspect of any theme sentence
  - A small dictionary of very basic domain independent sentiment words and a dictionary of negative words (created manually)
  - Sentences with exclamatory and question mark
  - A rule based lightweight anaphora resolution engine based on POS tags to handle sentences where subjects are replaced with anaphors

Subject-Aspect-Evaluation

- Opinion unit as a quadruple, i.e., opinion holder, subject, aspect, evaluation of an opinion.
- Opinion holder not identified in the present task
- Linguistic rules based on POS and chunk information for subject, aspect and evaluation identification
- Sentiment and negative words help to identify subject, aspect, and evaluation
- NPs, VPs and adjective or adverbial phrases play a crucial role

Results

- **Opinionated and non-opinionated sentences.**

Query	Precision	Recall
Q1	80.00%	83.33%
Q2	79.16%	76.00%
Q3	73.33%	78.57%
Q4	79.31%	76.66%
Q5	75.75%	71.42%
Overall	77.51%	77.19%

- **<Subject, Aspect, Evaluation> feature identification.**

Query	Precision	Recall
Q1	62.5%	66.66%
Q2	73.07%	61.29%
Q3	70.80%	60.02%
Q4	71.2%	58.78%
Q5	61.60%	55.66%
Overall	63.87%	60.48%