Topic-Based Bengali Opinion Summarization

Amitava Das
Department of Computer Science and Engineering
Jadavpur University
amitava.santu@gmail.com

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
sivaji_cse_ju@yahoo.com

Abstract

In this paper the development of an opinion summarization system that works on Bengali News corpus has been described. The system identifies the sentiment information in each document, aggregates them and represents the summary information in text. The present sys-tem follows a topic-sentiment model for sentiment identification and aggregation. Topic-sentiment model is designed as discourse level theme identification and the topic-sentiment aggregation is achieved by theme clustering (k-means) and Document level Theme Relational Graph representation. The Document Level Theme Relational Graph is finally used for candidate summary sentence selection by standard page rank algorithms used in Information Retrieval (IR). As Bengali is a resource constrained language, the building of annotated gold standard corpus and acquisition of linguistics tools for lexico-syntactic, syntactic and discourse level features extraction are described in this paper. The reported accuracy of the Theme detection technique is 83.60% (precision), 76.44% (recall) and 79.85% (F-measure). The summarization system has been evaluated with Precision of 72.15%, Recall of 67.32% and F-measure of 69.65%.

1 Introduction

The Web has become a rich source of various opinions in the form of product reviews, travel advice, social issue discussions, consumer complaints, movie review, stock market predictions, real estate market predictions, etc. Present computational systems need to extend the power of understanding the sentiment/opinion expressed in an electronic text to act properly in the society rather than dealing with the topic of a document. The topic-document model of information retrieval has been studied for a long time and systems are available publicly since last decade. On the contrary Opinion Mining/Sentiment Analysis is still an unsolved research problem. Although a few systems like Twitter Sentiment Analysis Tool[1], TweetFeel[2] are available in World Wide Web since last few years still more research efforts are necessary to match the user satisfaction level and social need.

Researchers have taken multiple approaches towards the problem of Opinion Summarization like Topic-sentiment model, Textual summaries at single document or multiple document perspective and graphical summaries or visualization. The works on opinion tracking systems have explicitly incorporated temporal dimension. The topic-sentiment model is well established for opinion retrieval.

The concept of reputation system was first introduced in (Resnick et al., 2000). Reputation systems for both buyers and sellers are needed to earn each other's trust in online interactions.

Ku et al., (2005) selects representative words from a document set to identify the main concepts in the document set. A term is considered to represent a topic if it appears frequently across documents or in each document. Different methodologies have been used to assign weights to each word both at document level and paragraph level. The precision and recall values of the system have been reported as 0.56 and 0.85.

[1] http://twittersentiment.appspot.com/
[2] http://www.tweetfeel.com/

Coling 2010: Poster Volume, pages 232–240,
Beijing, August 2010

Zhou et al. (2006) have proposed the architecture for generative summary from blogosphere. Typical multi-document summarization (MDS) systems focus on content selection followed by synthesis by removing redundancy across multiple input documents. The online discussion summarization system (Zhou et al., 2006) work on an online discussion corpus involving multiple participants and discussion topics are passed back and forth by various participants. MDS systems are insufficient in representing this aspect of the interactions. Due to the complex structure of the dialogue, similar subtopic structure identification in the participant-written dialogues is essential. Maximum Entropy Model (MEMM) and Support Vector Machine (SVM) have been used with a number of relevant features.

Carenini et al. (2006) present and compare two approaches to the task of multi document opinion summarization on evaluative texts. The first is a sentence extraction based approach while the second one is a natural language generation-based approach. Relevant extracted features are categorized in two types: User Defined Features (UDF) and Crude Features (CF) as described in (Hu and Liu, 2004).

The summary generation technique uses the aggregation of the extracted features, CF and UDF. Opinion aggregation has been done by the two relevant features: opinion strength and polarity. A new opinion distribution function feature has been introduced to capture the overall opinion distributed in corpus.

Kawai et al. (2007) developed a news portal site called Fair News Reader (FNR) that recommends news articles with different sentiments for a user in each of the topics in which the user is interested. FNR can detect various sentiments of news articles and determine the sentimental preferences of a user based on the sentiments of previously read articles by the user. News articles crawled from various news sites are stored in a database. The contents are integrated as needed and the summary is presented on one page. A sentiment vector on the basis of word lattice model has been generated for every document. A user sentiment model has been proposed based on user sentiment state. The user sentiment state model works on the browsing history of the user. The intersection of the documents under User Vector and Sentiment Vector are the results.

2 Resource Organization

Resource acquisition is one of the most challenging obstacles to work with resource constrained languages like Bengali. Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh. Extensive NLP research activities in Bengali have started recently but resources like annotated corpus, various linguistic tools are still unavailable for Bengali in the required measure. The manual annotation of gold standard corpus and acquisition of various tools used in the feature extraction for Bengali are described in this section.

2.1 Gold Standard Data Acquisition

2.1.1 Corpus

For the present task a Bengali news corpus has been developed from the archive of a leading Bengali news paper available on the Web (http://www.anandabazar.com/). A portion of the corpus from the editorial pages, i.e., Reader's opinion section or Letters to the Editor Section containing 28K word forms has been manually annotated with sentence level subjectivity and discourse level theme words. Detailed reports about this news corpus development in Bengali can be found in (Das and Bandyopadhyay, 2009b).

2.1.2 Annotation

From the collected document set (Letters to the Editor Section), some documents have been chosen for the annotation task. Some statistics about the Bengali news corpus is represented in the Table 1. Documents that have appeared within an interval of four months are chosen on the hypothesis that these letters to the editors will be on related events. A simple annotation tool has been designed for annotating the sentences considered to be important for opinion summarization. Three annotators (Mr. X, Mr. Y and Mr. Z) participated in the present task.

```
<Story>
……..
……..
<SS><TW>Sargeant O'Leary</TW> said "the
<TW>incident</TW> took place at 2:00pm."</SS>
……..
</Story>
```

Figure 1: XML Annotation Format

Annotators were asked to annotate sentences for summary and to mark the theme words (topical expressions) in those sentences. The documents with such annotated sentences are saved in

XML format. Figure 1 shows the XML annotation format. "<SS>" marker denotes subjective sentences and "<TW>" denotes the theme words.

Bengali NEWS Corpus Statistics	
Total number of documents in the corpus	100
Total number of sentences in the corpus	2234
Average number of sentences in a document	22
Total number of wordforms in the corpus	28807
Average number of wordforms in a document	288
Total number of distinct wordforms in the corpus	17176

Table 1: Bengali News Corpus Statistics

The annotation tool highlights the sentiment words (Das and Bandyopadhyay, 2010a)[3] by four different colors within a document according to their POS categories (Noun, Adjective, Adverb and Verb). This technique helps to increase the speed of annotation process. Finally 100 annotated documents have been developed.

2.1.3 Inter-annotator Agreement

The agreement of annotations among three annotators has been evaluated. The agreements of tag values at theme words level and sentence levels are listed in Tables 2 and 3 respectively.

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg
Percentage	82.64%	71.78%	80.47%	78.30%
All Agree	69.06%			

Table 2: Agreement of annotators at theme words level

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg
Percentage	73.87%	69.06%	60.44%	67.8%
All Agree	58.66%			

Table 3: Agreement of annotators at sentence level

From the analysis of inter-annotator agreement, it is observed that the agreement drops fast as the number of annotator's increases. It is less possible to have consistent annotations when more annotators are involved. In the present task the inter-annotator agreement is better for theme words annotation rather than candidate sentence identification for summary though a small number of documents have been considered.

Further discussion with annotators reveals that the psychology of annotators is to grasp as many as possible theme words identification during annotation but the same groups of annotators are more cautious during sentence identification for summary as they are very conscious to find out the most concise set of sentences that best describe the opinionated snapshot of any document. The annotators were working independent of each other and they were not trained linguists.

2.2 Subjectivity Classifier

Work in opinion mining and classification often assumes the incoming documents to be opinionated. Opinion mining system makes false hits while attempting to summarize non-subjective or factual sentences or documents. It becomes imperative to decide whether a given document contains subjective information or not as well as to identify which portions of the document are subjective or factual. This task is termed as subjectivity detection in sentiment literature. The subjectivity classifier that uses SVM machine learning technique and described in (Das and Bandyopadhyay, 2009a) has been used here. The recall measure of the present classifier is greater than its precision value. The evaluation results of the classifier are 72.16% (Precision) and 76.00 (recall) on the News Corpus.

2.3 Feature Organization

The set of features used in the present task have been categorized as Lexico-Syntactic, Syntactic and Discourse level features. These are listed in the Table 4 below and have been described in the subsequent subsections.

Types	Features
Lexico-Syntactic	POS
	SentiWordNet
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing Depth
Discourse Level	Title of the Document
	First Paragraph
	Term Distribution
	Collocation

Table 4: Features

2.3.1 Lexico-Syntactic Features

2.3.1.1 Part of Speech (POS)

It has been shown in (Hatzivassiloglou et. al., 2000), (Chesley et. al., 2006) etc. that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like (Nasukawa et. al., 2003) are mostly based on adjective words. Details of the Bengali POS tagger used can be found in (Das and Bandyopadhyay 2009b).

[3] http://www.amitavadas.com/sentiwordnet.php

2.3.1.2 SentiWordNet (Bengali)

Words that are present in the SentiWordNet carry opinion information. The developed SentiWordNet (Bengali) (Das and Bandyopadhyay, 2010a) is used as an important feature during the learning process. These features are individual sentiment words or word n-grams (multiword entities) with strength measure as strong subjective or weak subjective. Strong and weak subjective measures are treated as a binary feature in the supervised classifier. Words which are collected directly from SentiWordNet (Bengali) are tagged with positivity or negativity score. The subjectivity score of these words are calculated as:

$$E_s = | S_p | + | S_n |$$

where E_s is the resultant subjective measure and S_p , S_n are the positivity and negativity scores respectively.

2.3.1.3 Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document. The system generates four separate high frequent word lists for four POS categories: Adjective, Adverb, Verb and Noun after function words are removed. Word frequency values are then effectively used as a crucial feature in the Theme Detection technique.

2.3.1.4 Stemming

Several words in a sentence that carry opinion information may be present in inflected forms and stemming is necessary for them before they can be searched in appropriate lists. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer (Das and Bandyopadhyay, 2010b) based on stemming cluster technique has been used. This stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center.

2.3.2 Syntactic Features

2.3.2.1 Chunk Label

Chunk level information is effectively used as a feature in supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. In the task of identification of Theme expressions, chunk label markers play a crucial role. Further details of development of chunking system could be found in (Das and Bandyopadhyay 2009b).

2.3.2.2 Dependency Parser

Dependency depth feature is very useful to identify Theme expressions. A particular Theme word generally occurs within a particular range of depths in a dependency tree. Theme expressions may be a Named Entity (NE: person, organization or location names), a common noun (Ex: accident, bomb blast, strike etc) or words of other POS categories. It has been observed that depending upon the nature of Theme expressions it can occur within a certain depth in the dependency tree for the sentence. A statistical dependency parser has been used for Bengali as described in (Ghosh et al., 2009).

2.3.3 Discourse Level Features

2.3.3.1 Positional Aspect

Depending upon the position of the thematic clue, every document is divided into a number of zones. The features considered for each document are Title words of the document, the first paragraph words and the words from the last two sentences. A detailed study was done on the Bengali news corpus to identify the roles of the positional aspect features of a document (first paragraph, last two sentences) in the detection of theme words and subjective sentences for generating the summary of the document. The importance of these positional features is shown in Tables 5 on the Bengali gold standard set.

2.3.3.2 Title Words

Title words of a document always carry some meaningful thematic information. The title word feature has been used as a binary feature during CRF based machine learning.

2.3.3.3 First Paragraph Words

People usually give a brief idea of their beliefs and speculations in the first paragraph of the document and subsequently elaborate or support them with relevant reasoning or factual information. Hence first paragraph words are informative in the detection of Thematic Expressions.

2.3.3.4 Words From Last Two Sentences

Generally every document concludes with a summary of the opinions expressed in the document.

Positional Factors	Bengali
First Paragraph	56.80%
Last Two Sentences	78.00%

Table 5: Statistics on Positional Aspect.

2.3.3.5 Term Distribution Model

An alternative to the classical TF-IDF weighting mechanism of standard IR has been proposed as a model for the distribution of a word. The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. As discussed in Section 1, Carenini et al. (2006) have introduced the opinion distribution function feature to capture the overall opinion distributed in the corpus. Thus the objective is to estimate $f_d(w_i)$ that measures the distribution pattern of the k occurrences of the word w_i in a document d. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of topic-sentiment informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows:

$$f_d(w_i) = \sum_{i=1}^{n} (S_i - S_{i-1}) / n + \sum_{i=1}^{n} (TW_i - TW_{i-1}) / n$$

where n=number of sentences in a document with a particular theme word, S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

2.3.3.6 Collocation

Collocation with other thematic word/expression is undoubtedly an important clue for identification of theme sequence patterns in a document. A window size of 5 including the present word is considered during training to capture the collocation with other thematic words/expressions.

3 Theme Detection

Term Frequency (TF) plays a crucial role to identify document relevance in Topic-Based Information Retrieval. The motivation behind developing Theme detection technique is that in many documents relevant words may not occur frequently or irrelevant words may occur frequently. Moreover for sentiment analysis topic words should have sentiment conceptuality. The Theme detection technique has been proposed to resolve these issues to identify discourse level relevant topic-semantic nodes in terms of word or expressions using a standard machine learning technique. The machine learning technique used here is Conditional Random Field (CRF)[4]. The theme word detection is defined as a sequence labeling problem. Depending upon the series of input feature, each word is tagged as either Theme Word (TW) or Other (O).

4 Theme Clustering

Theme clustering algorithms partition a set of documents into finite number of topic based groups or clusters in terms of theme words/expressions. The task of document clustering is to create a reasonable set of clusters for a given set of documents. A reasonable cluster is defined as the one that maximizes the within-cluster document similarity and minimizes between-cluster similarities. There are two principal motivations for the use of this technique in theme clustering setting: efficiency, and the **cluster hypothesis**.

The **cluster hypothesis** (Jardine and van Rijsbergen, 1971) takes this argument a step further by asserting that retrieval from a clustered collection will not only be more efficient, but will in fact improve retrieval performance in terms of recall and precision. The basic notion behind this hypothesis is that by separating documents according to topic, relevant documents will be found together in the same cluster, and non-relevant documents will be avoided since they will reside in clusters that are not used for retrieval. Despite the plausibility of this hypothesis, there is only mixed experimental support for it. Results vary considerably based on the clus-

[4] http://crfpp.sourceforge.net

tering algorithm and document collection in use (Willett, 1988; Shaw et al., 1996).

Application of the clustering technique to the three sample documents results in the following theme-by-document matrix, A, where the rows represent Docl, Doc7 and Doc13 and the columns represent the themes politics, sport, and travel.

$$A = \begin{bmatrix} election & cricket & hotel \\ parliament & sachin & vacation \\ governor & soccer & tourist \end{bmatrix}$$

The similarity between vectors is calculated by assigning numerical weights to these words and then using the cosine similarity measure as specified in the following equation.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \vec{q}_k . \vec{d}_j = \sum_{i=1}^{N} w_{i,k} \times w_{i,j} \text{ ---- (1)}$$

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a similarity metric, since it is too sensitive to the absolute magnitudes of the various dimensions. However, the dot product between vectors that have been length normalized has a useful and intuitive interpretation: it computes the **cosine** of the angle between the two vectors. When two documents are identical they will receive a cosine of one; when they are orthogonal (share no common terms) they will receive a cosine of zero. Note that if for some reason the vectors are not stored in a normalized form, then the normalization can be incorporated directly into the similarity measure as follows.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \text{ ----(2)}$$

Of course, in situations where the document collection is relatively static, it makes sense to normalize the document vectors once and store them, rather than include the normalization in the similarity metric.

Calculating the similarity measure and using a predefined threshold value, documents are classified using standard bottom-up soft clustering k-means technique. The predefined threshold value is experimentally set to 0.5 as shown in Table 6.

A set of initial cluster centers is necessary in the beginning. Each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}$ (where $\vec{\mu}$ is the clustering coefficient) of its members, that is $\vec{\mu} = \left(1/\left|c_j\right|\right)\sum_{x \in c_j} \vec{x}$. The distance function is the **cosine vector** similarity function.

ID	Themes	1	2	3
1	**প্রশাসন** (administration)	**0.63**	0.12	0.04
1	সুশাসন (good-government)	**0.58**	0.11	0.06
1	সমাজ (Society)	**0.58**	0.12	0.03
1	আইন (Law)	**0.55**	0.14	0.08
2	গবেষণা (Research)	0.11	**0.59**	0.02
2	কলেজ (College)	0.15	**0.55**	0.01
2	**উচ্চশিক্ষা (Higher Study)**	0.12	**0.66**	0.01
3	জেহাদি (Jehadi)	0.13	0.05	**0.58**
3	মসজিদ (Mosque)	0.05	0.01	**0.86**
3	মুশারফ (Musharaf)	0.05	0.01	**0.86**
3	**কাশ্মীর (Kashmir)**	0.03	0.01	**0.93**
3	পাকিস্তান (Pakistan)	0.06	0.02	**0.82**
3	নয়াদিল্লী (New Delhi)	0.12	0.04	**0.65**
3	বর্ডার (Border)	0.08	0.03	**0.79**

Table 6: Five cluster centroids (mean $\vec{\mu}_j$)

Table 6 gives an example of theme centroids from the K-means clustering. Bold words in Theme column are cluster centers. Cluster centers are assigned by maximum clustering coefficient. For each theme word, the cluster from table 6 is still the dominating cluster. For example, "প্রশাসন" has a higher membership probability in cluster 1. But each theme word also has some non-zero membership in all other clusters. This is useful for assessing the strength of association between a theme word and a topic. Comparing two members of the cluster2, "কাশ্মীর" and "নয়াদিল্লী", it is seen that "নয়াদিল্লী" is strongly associated with cluster2 (p=0.65) but has some affinity with other clusters as well (e.g., p =0.12 with the cluster1). This is a good example of the utility of soft clustering. These non-zero values are still useful for calculating vertex weights during Theme Relational Graph generation.

5 Construction of Document Level Theme Relational Graph

Representation of input text document(s) in the form of graph is the key to our design principle. The idea is to build a document graph G=<V,E> from a given source document $d \in D$. First, the input document *d* is parsed and split into a number of text fragments (sentence) using sentence delimiters (Bengali sentence marker "।", "?" or "!"). At this preprocessing stage, text is tokenized, stop words are eliminated, and words are

stemmed (Das and Bandyopadhyay, 2010b). Thus, the text in each document is split into fragments and each fragment is represented with a vector of constituent theme words. These text fragments become the nodes V in the document graph.

The similarity between two nodes is expressed as the weight of each edge E of the document graph. A weighted edge is added to the document graph between two nodes if they either correspond to adjacent text fragments in the text or are semantically related by theme words. The weight of an edge denotes the degree of the relationship. The weighted edges not only denote document level similarity between nodes but also inter document level similarity between nodes. Thus to build a document graph G, only the edges with edge weight greater than some predefined threshold value are added to G, which basically constitute the edges E of the graph G.

The Cosine similarity measure has been used here. In cosine similarity, each document d is denoted by the vector $\vec{V}(d)$ derived from d, with each component in the vector for each Theme words. The cosine similarity between two documents (nodes) d1 and d2 is computed using their vector representations $\vec{V}(d1)$ and $\vec{V}(d2)$ as equation (1) and (2) (Described in Section 4). Only a slight change has been done i.e. the dot product of two vectors $\vec{V}(d1) \bullet \vec{V}(d2)$ is defined as $\sum_{i=1}^{M} V(d1)V(d2)$. The Euclidean length of d is defined to be $\sqrt{\sum_{i=1}^{M} \vec{V}_i^{\,2}(d)}$ where M is the total number of documents in the corpus. Theme nodes within a cluster are connected by vertex, weight is calculated by the clustering co-efficient of those theme nodes. No inter cluster vertex are there. Cluster centers are interconnected with weighted vertex. The weight is calculated by cluster distance as measured by cosine similarity measure as discussed earlier.

To better aid our understanding of the automatically determined category relationships we visualized this network using the Fruchterman-Reingold force directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL network analysis tool (Smith et al., 2009)[5]. A theme relational model graph drawn by NodeXL is shown in Figure 2.

[5] Available from http://www.codeplex.com/NodeXL

6 Summarization System

Present system is an extractive opinion summarization system for Bengali. In the previous sections, we described how to identify theme clusters that relates to different shared topics and subtopics, from a given input document set. But identifying those clusters is not only a step toward generating document level opinionated news summary rather another major step is to extract thematic sentences from each theme cluster that reflects the contextual concise content of the current theme cluster. Extraction of sentences based on their importance in representing the shared subtopic (cluster) is an important issue and it regulates the quality of the output summary. We have used Information Retrieval (IR) based technique to identify the most "informed" sentences from any cluster and it can be termed as IR based cluster center for that particular cluster. With the adaptation of ideas from page rank algorithms (Page et al., 1998), it can be easily observed that a text fragment (sentence) in a document is relevant if it is highly related to many relevant text fragments of other documents in the same cluster. Since, in our document graph structure, the edge score reflects the correlation measure between two nodes, it can be used to identify the most salient/informed sentence from a sentence cluster. We computed the relevance of a node/sentence by summing up the edge scores of those edges connecting the node with other nodes in the same cluster. Then the nodes are given rank according to their calculated relevance scores and the top ranking sentences is selected as the candidate sentence representing the opinion summary. For example four such candidate sentences are shown in Table 7. The words in bold are the theme words based on those theme words the sentences are extracted.

Candidate Sentence	IR Score
মহম্মদ আমিনের মতো **পলিটব্যুরোর** 'নবীনতম' সদস্যকেও কিন্তু বয়সের দিক হইতে নবীন ভাবা কঠিন।	151
এবার চিন্তা আরওএকটু বেশি, কারণ এই **মূল্যবৃদ্ধির** পিছনে যেমন দেশের ভিতরে জিনিসপত্রের জোগান কমে যাওয়া আছে, তেমনই আছে **আন্তর্জাতিক বাজারে মূল্যবৃদ্ধির** প্রবণতা।	167
স্বাধীনতার পর ষাট বছর গত হইল, এখনও প্রায় সকল **সরকারি পরিকল্পনার** পিছনে এই একটিই ভাবাদর্শ কাজ করে: বিভিন্ন **ভোটব্যাঙ্ককে** তুষ্ট করিয়া যেন তেন প্রকারেণ নিজেদের দলীয় স্থিতি নিশ্চিত করা।	130

Table 7: Candidate sentences

Another issue that is very important in summarization is sentence ordering so that the Output summary looks coherent. Once all the relevant sentences are extracted across the input documents, the summarizer has to decide in which order to present them so that the whole text makes sense for the user. We prefer the original order of sentences as they occurred in original document.

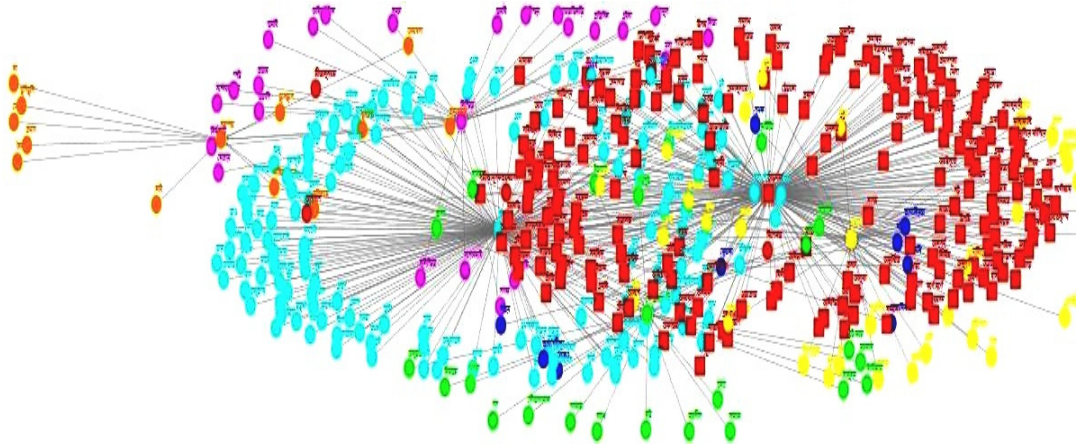

Figure 2: Document Level Theme Relational Graph by NodeXL

7 Experimental Result

The evaluation result of the CRF-based Theme Detection task for Bengali is presented in Table 8. The result is presented individually for every annotators and the overall result of the system.

	Metrics	X	Y	Z	Avg
Theme Detection	**Precision**	87.65%	85.06%	78.06%	83.60%
	Recall	80.78%	76.06%	72.46%	76.44%
	F-Score	84.07%	80.30%	75.16%	79.85%

Table 8: Results of CRF-based Theme Identifier

The evaluation result of subjective sentence identification of the system for opinion summary is in the Table 9.

	Metrics	X	Y	Z	Avg
Summarization	**Precision**	77.65%	67.22%	71.57%	72.15%
	Recall	68.76%	64.53%	68.68%	67.32%
	F-Score	72.94%	65.85%	70.10%	69.65%

Table 9: Final Results subjective sentence identification for summary

8 Error Analysis

The evaluation result of the present summarization system is reasonably good but still not outstanding. During the error analysis we found that the main false hits occurring for subjectivity identifier. It has been reported (Section 2.2) that the recall value of the classifier is higher than its precision. Hence some objective sentences are identified during subjectivity analysis. Some of the sentences get high score during Theme detection or Theme clustering and being included in final summary. Our observation is at least 2-3% sentences are included due to the wrong identification by Subjectivity identifier.

Another vital source of errors occurring in the accuracy level of linguistics resources and tools are the POS tagger, Chunker and Dependency Parser. These linguistics tools are not well performing hence the resultant Theme identification system is missing some of the important theme words. Successive Theme clustering, Document level weighted theme relational model fails to accumulate those important theme expressions. Our observation is at most 3-5% improvement could be possible on final system by granular improvement of every linguistic tool.

9 Conclusion

In this work we have reported our work on single-document opinion summarization for Bengali. The novelty of the proposed technique is the topic based document-level theme relational graphical representation. According to best of our knowledge this is the first attempt on opinion summarization for Bengali. The approach presented here is unique in every aspect as in literature and for a new language like Bengali.

Our next research target is to generate a hierarchical cluster of theme words with time-frame relations. Time-frame relations could be useful for time wise opinion tracking.

References

Carenini Giuseppe, Ng Raymond, and Pauls Adam. Multi-document summarization of evaluative text. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pages 305–312, 2006.

Chesley Paula, Vincent Bruce, Xu Li, and Srihari Rohini. Using verbs and adjectives to automatically classify blog sentiment. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 27–29, 2006.

Das, A. and Bandyopadhyay S. (2009b) Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII).

Das, A. and Bandyopadhyay, S. (2009a). Subjectivity Detection in English and Bengali: A CRF-based Approach., In Proceeding of ICON 2009, December 14th-17th, 2009, Hyderabad.

Das, A. and Bandyopadhyay, S. (2010a). SentiWordNet for Bangla. In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary , February 23th to 24th, 2010, Mysore.

Das, A. and Bandyopadhyay, S. (2010b). Morphologi-cal Stemming Cluster Identification for Bangla., In Knowledge Sharing Event-1: Task 3: Morphologi-cal Analyzers and Generators, January, 2010, Mysore.

Fruchterman Thomas M. J. and Edward M. Reingold. 1991. Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164.

Ghosh A., Das A., Bhaskar P., Bandyopadhyay S. (2009). Dependency Parser for Bengali : the JU System at ICON 2009., In NLP Tool Contest ICON 2009, December 14th-17th, 2009a, Hyderabad.

Hatzivassiloglou Vasileios and Wiebe Janyce. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 299-305, 2000.

Hu M. and Liu B.. 2004a. Mining and summarizing-customer reviews. In Proc. of the 10th ACM-SIGKDD Conf., pages 168–177, New York, NY, USA. ACM Press.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. In-formation Storage and Retrieval, 7, 217-240.

Kawai Yukiko, Kumamoto Tadahiko, and Katsumi Tanaka. Fair News Reader: Recommending news articles with different sentiments based on user pre-ference. In Proceedings of Knowledge-Based Intel-ligent Information and Engineering Systems (KES), number 4692 in Lecture Notes in Computer Science, pages 612–622, 2007.

Ku Lun-Wei, Li Li-Ying, Wu Tung-Ho, and Chen Hsin-Hsi. Major topic detection and its application to opinion summarization. In Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), pages 627–628, 2005. Poster paper.

Nasukawa Tetsuya and Yi Jeonghee. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), pages 70-77, 2003.

Page Lawrence, Brin Sergey, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

Resnick Paul, Kuwabara Ko, Zeckhauser Richard, and Friedman Eric. Reputation systems. Communications of the Association for Computing Machinery (CACM), 43(12):45–48, 2000. ISSN 0001-0782.

Smith Marc, Shneiderman Ben, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. 2009. Analyzing (social media) networks with NodeXL. In C&T '09: Proc. Fourth International Conference on Communities and Technologies, Lecture Notes in Computer Science. Springer.

Willerr, P. (1988). Recent trends in hierarchic document clustering: A critical review. Information Processing and Management, 24(5), 577-597.

Zhou Liang and Hovy Eduard. On the summarization of dynamically introduced information: Online dis-cussions and blogs. In AAAI Symposium on Com-putational Approaches to Analysing Weblogs (AAAI-CAAW), pages 237–242, 2006.