Finding Emotion Holder from Bengali Blog Texts –An Unsupervised Syntactic Approach*

Dipankar Das and Sivaji Bandyopadhyay

Department of Computer Science and Engineering, Jadavpur University,
188, Raja S.C. Mullick Road, Kolkata -700032, West Bengal
dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

Abstract. This paper presents two different approaches for identifying *emotion holder*s from Bengali blog sentences. Two types of strategies yield average agreement measures of *0.78* and *0.80* for annotating *emotion holder*s with respect to all emotion classes. The baseline model is developed based on the combinations of various part-of-speech (*POS*) features extracted from the phrase-based similarities. The syntactic model is based on the argument structure of the sentences with respect to the verbs. If the acquired argument structure of a Bengali blog sentence with respect to its verb matches with any of the frame syntax retrieved for its equivalent English verb of identical sense from VerbNet, the holder role associated with the English VerbnNet frame is mapped to the appropriate slot in Bengali sentence. The syntactic model with an average *F-score* of 60.03% outperforms the baseline model with an average *F-score* of 50.85% on 500 test sentences.

Keywords: Emotion Holder, Kappa, Argument Structure, Syntax, VerbNet.

1 Introduction

Researches on *emotion holder* extraction are important for discriminating emotions that are viewed from different perspectives (Seki, 2007). *Emotion holder* inscribed in natural language texts plays an important role with respect to the reader or writer. A wide range of natural language processing (NLP) tasks such as tracking users' emotion about reviews or events or politics as expressed in online forums or news, customer relationship management all are using the emotional information.

Blogs on the other hand are the communicative and informative repository of text based emotional contents in the Web 2.0 (Lin *et al.*, 2007). Sometimes, blog posts are annotated by other bloggers. The utilization of blog medium containing users' emotional contents is therefore considered as an affective substrate to analyze the reaction of emotion catalyst like *emotion holder*.

In the present task, identification of *emotion holder* is attempted for Bengali; a less privileged, less computerized and morphologically rich language. There is no existing *emotion holder* annotated corpus in Bengali. Manual annotation of *emotion holder* and the successive inter-annotator agreements have been carried out on a small set of 500 sentences of the Bengali blog corpus (Das and Bandyopadhyay, 2009). The corpus is tagged with Ekman's (1993) six emotion types at sentence level. The phrase based similarity clues containing different part-of-speech (*POS*) combinations of the blog sentences are considered as the probable candidates of *emotion holder* for the baseline model.

* The work reported in this paper was supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

Developing syntactic model based on argument structure satisfies the demands of the baseline model as well. The pivotal hypothesis considered in the syntactic model is based on the hypothesis followed in (Das *et al.*, 2009; Banerjee *et al.*, 2009). The verb-based argument structures are acquired from the POS tagged and chunked Bengali blog corpus. Equivalent English verbs of identical sense for the Bengali verbs are extracted using Bengali to English bilingual dictionary[1]. The available frames of the equivalent English verbs are retrieved from English VerbNet (Kipper-Schuler, 2005). If an acquired argument structure of a sentence matches with any of its equivalent retrieved frames, the holder role (e.g *Experiencer, Agent, Actor, Beneficiary* etc.) associated with the English VerbNet frame is mapped to the appropriate slot in Bengali sentence considering its acquired argument structure. The *F-scores* of the baseline and syntax based models are 50.85% and 60.03% on 500 test sentences respectively.

The rest of the paper is organized as follows. Section 2 describes the related work. The preparation of the annotated corpus and the baseline model are described in Section 3 and Section 4 respectively. Development of syntactic model is discussed in Section 5. Evaluation mechanism along with the associated results is mentioned in Section 6. Finally Section 7 concludes the paper.

2 Related Work

The work on labeling the arguments of the verbs with their semantic roles using a novel frame matching technique is mentioned in (Swier and Stevenson, 2004). Identification of the opinion propositions and their holders is described in (Bethard *et al.*, 2004) mainly for verbs. Identification of opinion holders for Question Answering with supporting annotation task has been attempted from the very beginning (Wiebe *et al.*, 2005). (Choi *et al.*, 2005) used the named entities (*NEs*) to identify the opinion holders with the help of machine learning and pattern-based techniques. Based on the traditional perspectives, another work discussed in (Hu *et al.*, 2006) uses an emotion knowledge base for extracting *emotion holder*. The machine learning based classification task for "not holder", "weak holder", "medium holder", or "strong holder" is carried out in (Evans, 2007). Kim and Hovy (2006) identified opinion holder with topic from media text using semantic role labeling. An anaphor resolution based opinion holder identification method exploiting lexical and syntactic information from online news documents is carried out in (Kim *et al.*, 2007). The syntactic models of identifying *emotion holder* for English emotional verbs are discussed in (Das and Bandyopadhyay, 2010).

The above works are closely related to the present one. But the present approach aims to acquire all probable *emotion holders* from a sentence if there multiple occurrences exist. Apart from utilizing traditional hints like named entities or anaphors, the present syntactic similarity based holder identification technique performs satisfactorily in Bengali. Moreover, all the above-cited works have been attempted for English. Recent study shows that non-native English speakers support the growing use of the Internet[2]. This raises the demand of linguistic resources for languages other than English. Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh but it is less computerized compared to English. To the best of our knowledge, at present, there is no such prior work of *emotion holder* identification has been conducted for Indian languages. Thus we believe that the present study of identifying *emotion holder* would help in the development of emotion analysis systems as well.

3 Emotion Holder (EH) Annotation

The source or holder of an emotional expression is the speaker or writer or experiencer. The main criteria considered for annotating *emotion holder*s are based on the nested source hypothesis as described in (Wiebe *et al.*, 2005). The structure of Bengali blog corpus (as shown in Figure 1)

[1] http://home.uchicago.edu/~cbs2/banglainstruction.html
[2] http://www.internetworldstats.com/stats.htm

helps in the holder annotation process. Sometimes, the comments of one blogger are annotated by other bloggers in the blog posts. Thus the holder annotation task in user comments sections was less cumbersome than annotating the holders inscribed in the topic section.

```
-<DOC docid = xyz>
    -<Topic>…. </Topic>
    -<User Comments>
        -<U uid=1>… </U>
        -<U uid=2>… </U>
        -<U uid=3>….
            -<U uid=1>…</U>
…</U>…
    </User Comments>
</DOC>
```

Figure 1: General structure of a blog document.

Prior work in identification of opinion holders has sometimes identified only a single opinion per sentence (Bethard *et al.*, 2004), and sometimes several (Choi *et al.*, 2005). As the blog corpus has sentence level emotion annotation, the former category is adopted. But, it is observed that the long sentences contain more than one emotional expression and hence associated with multiple *emotion holder*s (EH). All probable *emotion holder*s of a sentence are stored in an anchoring vector. If multiple *emotion holder*s exist, the successive holders are annotated and placed in the vector according to their order of occurrences.

The annotation of *emotion holder* at sentence level requires the knowledge of two basic constraints (*explicit* and *implicit*) separately. The *explicit* constraints qualify single prominent *emotion holder* that is directly involved with the emotional expression whereas the *implicit* constraints qualify all direct and indirect nested sources as *emotion holder*s. For example, in the following Bengali sentences, the pattern shown in **bold** face denotes the *emotion holder*. In the second example, the appositive case (e.g. **রামের** সুখ (*Ram's pleasure*)) is also identified and placed in the vector by removing the inflectional suffix (-**এর** in this case). Example 2 and Example 3 contain the *emotion holder*s ***রাম*** (*Ram*) and ***নাসরিন সুলতানা*** (*Nasreen Sultana*) based on *implicit* constraints.

Example 1. *EH_Vector*: < **সায়ণ** >
সায়ণ ভীষণ আনন্দ অনুভব করেছিল
(**Sayan**) (bhishon) (anondo) (anubhob) (korechilo)
Sayan felt very happy.

Example 2. *EH_Vector*: < **রাশেদ** , ***রাম*** >
রাশেদ অনুভব করেছিল যে **রামের** সুখ অন্তহীন
(*Rashed*) (anubhob) (korechilo) (je) (**Ramer**) (sukh) (antohin)
Rashed felt that Ram's pleasure is endless.

Example 3. *EH_Vector*: < **গেদু চাচা** , ***নাসরিন সুলতানা*** >
গেদু চাচা বলে : না গো বোন, আমি **নাসরিন সুলতানার** দুঃখের কথাতে
(Gedu ChaCha) (bole) : (na) (go) (bon) , (ami) (**Nasreen Sultanar**) (dookher) (kathate)
কেঁদে ফেলি।
(kende) (feli)
Gedu Chacha says: No my sister, I fall into cry on the sad speech of Nasreen Sultana

3.1 Agreement of Emotion Holder Annotation

The *emotion holder*s containing multi word Named Entities (*NEs*) are assumed as single *emotion holder* entities. As there is no agreement discrepancy in selecting the boundary of the single or multiple *emotion holder*s, we have used the standard metric, Cohen's *kappa* (κ) for measuring the inter-annotator agreement. Each of the elementary *emotion holder*s in an anchoring vector is treated as a separate *emotion holder* and the agreement between two annotators is carried out on each separate entity.

It is to be mentioned that the anchoring vectors provided by the two annotators may be disjoint. To emphasize the fact, additionally a simple technique is employed to measure the annotation agreement. If **X** is a set of *emotion holder*s selected by the first annotator and **Y** is a set of *emotion holder*s selected by the second annotator for an emotional sentence containing multiple *emotion holder*s, inter-annotator agreement **IAA** for that sentence is equal to quotient of number of *emotion holder*s in **X** and **Y** intersection divided by number of *emotion holder*s in **X** and **Y** union:

$$\mathbf{IAA = X \cap Y / X \cup Y}$$

Two types of agreement results per emotion class for annotating *emotion holder*s (EH) are shown in Table 1. Both types of agreements have been found satisfactory and the difference between the two agreement types is significantly less. The small difference indicates the minimal error involved in the annotation process. It is found that the agreement is highly moderate in case of single *emotion holder*, but is less in case of multiple holders. The disagreement occurs mostly in the case of satisfying the implicit constrains but some issues are resolved by mutual understanding.

Table 1: Kappa (κ) and [IAA] Agreements for Emotion Holder Annotation.

	A1-A2	A2-A3	A1-A3	Avg.
Happy [94, 118]	(0.87) [0.88]	(0.79) [0.81]	(0.76) [0.77]	(0.80) [0.82]
Sad [86, 92]	(0.82) [0.81]	(0.85) [0.83]	(0.78) [0.80]	(0.81) [0.81]
Anger [82,85]	(0.80) [0.79]	(0.75) [0.73]	(0.74) [0.71]	(0.76) [0.74]
Disgust [76, 87]	(0.70) [0.68]	(0.72) [0.69]	(0.83) [0.84]	(0.75) [0.73]
Fear [75, 96]	(0.85) [0.82]	(0.78) [0.77]	(0.79) [0.81]	(0.80) [0.80]
Surprise [87, 115]	(0.78) [0.80]	(0.81) [0.79]	(0.85) [0.83]	(0.81) [0.80]
Total [500, 593]	(0.79) [0.81]	(0.78) [0.80]	(0.77) [0.79]	(0.78) [0.80]

4 Baseline Model

The baseline model is developed based on the phrasal pattern containing similarity clues. The patterns are grouped according to part-of-speech (POS) categories. It is observed that the hints are present mostly in the user comment portions of the Bengali blogs. Each of the user comment portions started with the corresponding username is the default hint that helps in capturing the first holder present in the anchoring vector of nested sources.

The test sentences are passed through a Bengali part of speech tagger (Ekbal and Bandyopadhyay, 2008) based on Support Vector Machine (SVM) technique. The POS tagger

was developed with a tagset of 26 POS tags[2], defined for the Indian languages. The POS tagger has demonstrated an overall accuracy of approximately 90%. The POS tagged sentences contain the similarity pattern at lexical level.

The named entities that are tagged with NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun) and NN (Common noun) or PRP (Pronoun) and present at the beginning of a sentence point to the presence of the *emotion holder*. The similarity pattern consists of two phrasal constituents, the subject and the verb. The common portion containing the additional constituents is basically the floating part. As Bengali is a free phrase order language, the ordering between the verb and the floating portion is not fixed. But, the general pattern such as [<NNP/NNPC/NN/NNC/PRP> {<VBZ/VM><Common_Portion>}] is considered as the phrasal pattern for capturing the clue of an *emotion holder*. The components of the common parts are assembled as started by the hint given by the first occurring POS tags of types NNP or NNC or PRP in the tagged sentence. Reaching of the verb POS like VBZ or VM stops the incorporation of the components of the common portion. The rest of any component after the verb is therefore added to build the common portion. The similarity patterns exist mostly in the simple sentences. The complex or compound sentences are hard to classify into this category.

The utilization of chunked information helps in identifying the phrasal similarities of the baseline system in terms of *F-score*. Not only the chunk level information helps in improving the baseline system but also helps in capturing the floating syntax from the freely ordered chunks. The baseline system achieves a low average *F-score* of 50.85% on 500 test sentences. But, the baseline system fails to identify the nested emotion holders.

5 Syntactic Model

The syntactic way of identifying argument structures of the sentences and capturing *emotion holders* from the viewpoint of *thematic role* has therefore been considered as a favored way to meet up the demands of the baseline model. More specifically, the argument structure or subcategorization information for a verb plays a crucial role in identifying the *emotion holder* from a sentence. A subcategorization frame is a statement of what types of syntactic arguments a verb (or an adjective) takes, such as objects, infinitives, that-clauses, participial clauses, and subcategorized prepositional phrases (Manning, 1993).

The hypothesis that was considered in (Banerjee *et al.*, 2009) for extracting Bengali subcategorization frames is also considered in the present task. The hypothesis is that the verb subcategorization frames or argument structure for the equivalent English verbs (sharing the same sense) of a Bengali verb are the initial set of valid verb subcategorization frames for that Bengali verb. Irrespective of ordering, the phrase level similarities between Bengali and the English language helps in acquiring the argument structure.

5.1 Syntax Acquisition Framework

The argument structure of a Bengali emotional sentence contains phrase level head information that in turn conveys the sentential syntax information. As, there is no full-fledged parser available in Bengali, a rule-based chunker is used to chunk the POS tagged emotional sentences with an overall accuracy of 89.4%. To identify the verbs from the POS tagged and chunked corpus, the words that are tagged as main verb (VM) and belong to the verb group chunk (VGNF) in the corpus are identified. The lexical pattern identifies simple verbs. For the compound or conjunct verbs, the pattern such as {[XXX] (NN) [YYY] (VM)} are retrieved from the Bengali POS tagged and chunked corpus (e.g. [VGNF ((*ananda*(NN) *kara*(VM))] means *enjoy* etc.). The light verbs YYY generally occur in inflected forms. Different suffixes may be attached to a light verb

[2] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

depending on the various features such as Tense, Aspect, and Person. A Bengali stemmer with an accuracy of 97.09% uses a suffix list to identify the stem form of the retrieved Bengali verbs. Another table stores the stem form and the corresponding root form.

The determination of equivalent English verbs of a Bengali verb is carried out using a Bengali to English bilingual dictionary[3]. The method to extract the English equivalent synsets of the Bengali verbs is based on the work done by (Banerjee *et al.*, 2010). It is found that each of the English equivalent synsets occurs in each separate class of English VerbNet (Kipper-Schuler, 2005). VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information such as *thematic roles* and *semantic predicates*, with syntactic frames and *selectional restrictions*. Member verbs in the same VerbNet class share common syntactic frames, and thus they are believed to have the same syntactic behavior. The VerbNet files containing member verbs and possible subcategorization frames are stored in XML file format. Hence, the XML files are pre-processed to build up a general list that contains all verbs, their classes and possible subcategorization frames (primary as well as secondary). This pre-processed list is searched to retrieve the subcategorization frames for each verb of the English equivalent synsets corresponding to the Bengali verb. These acquired subcategorization frames are believed to be the valid set of argument structures for the concerned Bengali verbs (Das *et al.* 2009; Banerjee *et al.*, 2009).

On the other hand, the chunked Bengali sentences are passed through a rule based *phrasal-head* extraction module to identify the phrase level argument structure of the sentences corresponding to the position of the verbs. The extracted *head part* of every phrase from the chunked data is considered as the component part of the whole argument structure for a sentence. For example, in simple sentences the occurrence of the NNPC, NNP, NNC or NN tags preceded by the PRP (Pronoun) NNP, NNC, NN or NNPC tags (may contain case markers (e.g. কে [*ke*]) and followed by a verb gives similar frame syntax for "Basic Transitive" frame of the VerbNet.

রাম সিতাকে ভালবাসে
Ram Sitake bhalobase.
Ram loves Sita
Acquired Argument Structure: [NNP NNP-*ke* VM]
Simplified Extracted VerbNet Frame Syntax: [<NP value="*Experiencer*" ></VERB><NP-*theme*>]

Similarly, the argument structure and extracted sentential complement, "S" frame for the Example 2 of Section 3 is as follows,

Acquired Argument Structure: [NNP VM DET-*je* S]
Simplified Extracted VerbNet Frame Syntax: [<NP value="*Experiencer*" ></VERB>< S-*that* (Sentential –*that* Complement)>]

The following result is for Example 1 of Section 3.

Acquired Argument Structure: [NNP NN VM]
Simplified Extracted VerbNet Frame Syntax: [<NP value="*Experiencer*" ></VERB>< NP-*theme*)>]

There are examples for which the case markers in Bengali are required to identify the *emotion holder*s. The case markers are the useful hints to capture the selectional restrictions and play the key role in distinguishing the holders from other valid alternatives. If the acquired argument structure for a Bengali emotional sentence is matched with any of the retrieved frames of English VerbNet, the holder information (e.g *Experiencer, Agent, Actor, Beneficiary* etc.) associated with

[3] http://home.uchicago.edu/~cbs2/banglainstruction.html

the English frame syntax is mapped to the appropriate slot of the acquired Bengali argument structure. Tag conversion routines are developed to transform the POS of the system-generated argument structures into the POS of the VerbNet frames.

6 Evaluation

The evaluation of the baseline and syntax based models are carried out on the annotated test set with 500 sentences. The baseline system suffers in disambiguating the emotional holders for complex and compound sentences, as no full-fledged dependency parser is available in Bengali. Moreover, the free phrase order characteristics of Bengali make the holder acquisition task difficult. It is found that the baseline model fails to identify the holders specified using implicit hints (Example 2 of section 3). The error analysis suggests that the rich morphology and free phrase order nature of Bengali restricts the baseline model to capture the holder information. Hence, we have further explored the baseline task based on the shallow dependency based chunked results. The holder identification based on dependency results improves the baseline system as it compares the similarity from phrasal heads rather than POS tags.

Argument structure acquisition from chunked data of the shallow parser contributes effectively in holder identification task. The emotional sentences containing passive sense are often confusing and hence require the inclusion of an alternation strategy for minimizing the error. But, it is observed that the syntax-based model outperforms the baseline significantly. Distinguishing the arguments from the adjuncts and the holder identification for passive sentences are not handled in the baseline model. Moreover, it is to be mentioned that acquisition of syntax from less structured blog sentences produce an average *F-score* of 60.95% with respect to all emotion classes. The evaluation results of the baseline and syntax based systems for single and multiple *emotion holder*s on 500 test sentences are presented in Table 2.

Table 2: F-scores of six emotion classes for Baseline and Syntax-based models.

	Baseline Single	Baseline Multiple	Syntax Single	Syntax Multiple
Happy [94, 118]	.5367	.5106	.6414	.6122
Sad [86, 92]	.5096	.4790	.6035	.5988
Anger [82,85]	.5323	.5021	.6276	.6012
Disgust [76, 87]	.5014	.4983	.5708	.5535
Fear [75, 96]	.5187	.4921	.6113	.5944
Surprise [87, 115]	.5204	.5011	.6026	.5867
Total [500, 593]	**.5198**	**.4972**	**.6095**	**.5911**

7 Conclusion

The paper describes the *emotion holder* identification in Bengali based on the argument structure. The evaluation result suggests for incorporating dependency parsing to improve the *F-score* along with the handling of inflections to extract more frames. The anaphor resolution technique is the future task to accomplish the task of identifying implicit emotion holders from texts.

References

A. Ekbal and S. Bandyopadhyay. 2008. Web-based Bengali News Corpus for Lexicon Development and POS Tagging. *POLIBITS*, 37(2008), pp. 20-29, Mexico.

Bethard Steven., Yu H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D. 2004. Automatic Extraction of Opinion Propositions and their Holders. *In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.*

Choi, Y., Cardie, C., Riloff, E., Patwardhan, S. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. *In Proceedings of HLT/EMNLP.*

D.Das, A.Ekbal, and S.Bandyopadhyay. 2009. Acquiring Verb Subcategorization Frames in Bengali from Corpora. *ICCPOL-09*, LNAI-5459, pp.386-393, Hong Kong.

D Das and S. Bandyopadhyay. 2009. Word to Sentence Level Emotion Tagging for Bengali Blogs. *ACL-IJCNLP 2009*, pp. 149-152, Suntec, Singapore.

D Das and S. Bandyopadhyay. 2010. Emotion Holder for Emotional Verbs – The role of Subject and Syntax. *CICLing- 2010*, A. Gelbukh (Ed.), LNCS 6008, pp. 385–393, Romania.

Ekman, P. 1993. Facial expression and emotion. *American Psychologist*, 48(4), pp.384–392

Evans, D.K.2007. A low-resources approach to Opinion Analysis: Machine Learning and Simple Approaches. *NTCIR*

Hu, J., Guan, C., Wang, M., Lin, F.2006. Model of Emotional Holder. *In Shi, Z.-Z., Sadananda, R. (eds.) PRIMA 2006. LNCS (LNAI)*, vol. 4088, pp. 534–539. Springer, Heidelberg

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the Sources and Targets of Subjective Expressions. *LREC 2008.*

K. H.-Y. Lin, C. Yang and H.-H. Chen. 2007. What Emotions News Articles Trigger in Their Readers?. *SIGIR*, pp. 733-734.

Kipper-Schuler, K. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, PA.

Kim, S.-M., Hovy, E. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. *ACL* .

Kim, Y., Jung, Y., Myaeng, S.-H. 2007. Identifying Opinion Holders in Opinion Text from Online Newspapers. *In 2007 IEEE International Conference on Granular Computing*, pp. 699–702, doi:10.1109/GrC.2007.45

Manning, C.D.1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *In 31st Meeting of the ACL*, Columbus, Ohio, pp. 235–242.

Seki, Y.2007. Opinion Holder Extraction from Author and Authority Viewpoints. *In SIGIR 2007. ACM*, New York, 978-1-59593-597-7/07/0007.

S.Banerjee, D.Das and S.Bandyopadhyay. 2009. Bengali Verb Subcategorization Frame Acquisition - A Baseline Model. *ACL-IJCNLP-2009, ALR-7 Workshop,*pp. 76-83, Suntec, Singapore.

S.Banerjee, D.Das and S.Bandyopadhyay. 2010. Classification of Verbs – Towards Developing a Bengali Verb Subcategorization Lexicon. *GWC-2010*, pp. 76-83 , Mumbai, India.

Swier, R.S., Stevenson, S. 2004. Unsupervised Semantic Role Labelling. *In Proceedings of EMNLP.*

Wiebe, J., Wilson, T., Cardie, C. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation 1(2).*