Detection of Semantic Errors from Simple Bangla Sentences

K. M. Azharul Hasan, Muhammad Hozaifa
Computer Science and Engineering Department
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh.
azhasan@gmail.com, hozaifa.moaj@gmail.com

Sanjoy Dutta
Computer Science and Engineering Department
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh.
dsanjoy58@live.com

Abstract—**We describe a methodology to detect semantic errors from Bangla sentences. According to Bangla grammar, a single verb can have many forms depending on its tense and person of its subject. The subject of a sentence can be noun or pronoun, may indicate human, animal, or any non-living entity. There is a fixed semantic relation between every verb and subject and object of a sentence. For example, a non-living entity can never feel hungry but living entity feels. This semantic difference checking for its correctness in a language is very important for the purpose of machine learning study and intelligent agent development for human computer interaction. Semantic error detection for Bangla language is an important research problem because of the variety that Bangla language offers in its grammatical, structural and semantic diversity. In this paper, we have established the relationship between subject and verb as well as object and verb of Bangla sentence. Hence we have proposed an algorithm for semantic correctness of simple Bangle sentences. The algorithm can easily be extended for other forms such as complex and compound sentences**

Keywords—Bangla Language Processing, Bangla Grammer, Semantic analysis, Bangla Simple structure.

I. INTRODUCTION

The word semantics expresses a range of ideas, from the language to the highly technical. It is often used in ordinary language for denoting a problem of understanding that comes down to word selection or connotation[1]. In linguistics, semantics is that branch of linguistics that deals with the study of meaning, changes in meaning, and the principles that govern the relationship between sentences or words and their meanings [2]. Sounds, facial expressions, body languages also have semantic (meaningful) content and comprises several branches of study. So study of semantic detection of language enables the communities to interpret and perform specific tasks if the orders, sentence or symbols are meaningful. Semantic error detection is a challenging task for resource constrained language like Bangla. But The research on semantic correctness checking is very important for the purpose of machine learning, opinion mining and intelligent agent development for human computer interaction. There have been some of researches on Bangla language like Bangla grammar detection[3]-[5], opinion mining or sentiment detection from Bangla text[6][7], Bangla Character recognition[8] development, English to Bangla and Bangla to English Translation[9] and Bangla Text to Speech and Speech to Text Synthesis[10][11]. But the research on Semantic detection from Bangla text is still inferior because of the absence of standard corpus of Bangla words[12]. The semantic of a Bangla sentence basically depends on the verb(s) used in the sentence. In this paper, we consider simple sentences having single verb of the form Subject +Object + Verb (SOV). In the SOV form, the relation of the verb with subject and object is of two fold;

1. Whether the verb with the subject has a well formed structure with semantic compatibility (SV relation)
2. Whether the Object and Verb (OV relation) has semantic compatibility.

To establish these two relations, we have created classification table of Bangla verbs on the basis of tense, person and also a classification table of noun on the basis of person, species, gender etc. Hence we have established the subject verb relation and Object Verb relation to check the semantic correctness of both SV and OV relation.

II. SEMANTIC ERROR DETECTION FROM SIMPLE BANGLA SENTENCES

Detecting semantic errors difficult to treat and needs a lot of preprocessing works such as constructing semantic knowledge base and automatic error-detection algorithm based on this knowledge base. To construct this semantic knowledge base, we have formulated the problem in two main parts namely categorization and Relationship Validation and Acceptance Checking.

A. Categorization

Categorization implies that objects are grouped into categories, usually for some specific purpose. Categorization is fundamental in language, prediction, inference, decision making and in all kinds of environmental interaction [13]. We have categorized the words so that each of the words in the same category posses same semantic relationship with other entity. We chose a Bengali sentence consisting of complete basic structure of the form Subject + Object + Verb. For example:

মানুষ (sub) ভাত (object) খায় (verb)।

সে (sub) ভাত (object) খাবে (verb) ।

কাক (sub) আকাশে (object) উড়ে(verb) ।

These sentences are broken into subject, object and verb parts for the purpose of categorization. We have prepared a categorization table for the purpose of identifying class of verbs, class of subjects from living being and their inter-relationships based on social impacts and use of them. Table1 and Table 2 show some examples of categorization for nouns and verbs for Bangla.

TABLE I. TABLE OF NOUN CATEGORIZATION

Category	Members
মানুষ	মানুষ রহিম করিম রবি আমি তুমি সে
গরু	গরু, গাভী
পাখি	পাখি কাক বাবুই চড়ুই
মাছ	মাছ ইলিশ রুই কাতলা
বাঘ	বাঘ বাঘিনী

TABLE II. TABLE OF VERB CATEGORIZATION

Category	Members
উড়ে	উড়া উড়েছিল উড়বে উড়ল উড়ি উড়ে
ভাসে	ভাসছিল ভাসবে ভাসল ভাসি ভাসে
খায়	খাবে খেল খাইবে খেয়েছিল খাই খায়
গায়	গায় গাইবে গাচ্ছিল গাইবে গাইবে
চালায়	চালায় চলে চলছে চলবে

B. Relationship Validation and Acceptance Checking

From noun and verb categorization we categorize the subject and verb into a more general class and hence we develop the relationship. Based on the categorization of verbs and nouns, we developed a relationship for subject and verb as well as verb and object of the sentence.

Definition 1 (**SV relation):** If there is a well-established semantic bond between a subject (noun) of the sentence and verb then there is a true SV relation between the subject and verb otherwise the relation is false. For example "পাখি উড়ে" has a true SV relation and "মাছ উড়ে" has false SV relation.

Definition 2 (**OV relation):** If there is a well-established semantic bond between an Object of the sentence and verb then there is a true OV relation between the object and verb otherwise the relation is false. OV relation is true when there is a true SV relation in that sentence. For example "মানুষ ভাত খায়**"** has a true OV relation and "মানুষ ঘাস খায়" has false OV relation.

Using the SV relationship, we have created a Validation Table (VT) to check the semantic acceptance. The entries of VT are a Boolean relationship True (T) or False (F). If there is false SV relation between subject and verb then the entry is F otherwise the entry is T. If the entry is T then it has one more entry which indicates the OV relation because OV relation is established if there is a true SV relation. If the VT entry is true then the corresponding OV indicates a set for which the SV is true.

TABLE III. VALIDATION TABLE OF BOOLEAN RELATIONSHIP BETWEEN SUBJECT AND VERB

Verb	Subject				
	মানুষ	গরু	পাখি	মাছ	বাঘ
খায়	T/S11	T/S12	T/S13	T/S14	T/S15
উড়ে	F	F	T/S23	F	F
ভাসে	F	F	F	T/S34	F
সাঁতরায়	T/S41	T/S42	T/S43	T/S44	T/S45
গায়	T/S51	F	T/S53	F	F
করে	T/S61	F	T/S63	T/S64	T/S65
কাটে	T/S71	T/S72	F	F	F
পড়ে	T/S81	F	F	F	F
দেখে	T/S91	T/S92	T/S93	T/S94	T/S95
জ্বালায়	T/S101	F	F	F	F
খেলে	T/S111	F	F	F	F
বলে	T/S121	F	F	F	F
চালায়	T/S131	F	F	F	F
হয়	T/S141	F	F	F	T/S145
রাখে	T/S151	F	F	F	F
শিখে	T/S161	F	F	F	F

Table 3 shows a sample VT and Table 4 shows a sample SV set. We check the semantic acceptance of a sentence by checking whether there is a true SV relation; i.e. if the corresponding entry in VT is T. If there is a valid relationship between subject and verb then the set on the OV table is checked and if the object is a member of the set then the sentence is semantically correct otherwise incorrect.

TABLE IV. TABLE OF VERB CATEGORIZATION

S11	{গরু পাখি ফল ভাত মুরগী... etc }
S12	{ঘাস ভাত... etc}
S13	{পোকামাকড় কেঁচো ... etc}
. . .	. . .
S161	{কোরআন,গান ... etc}

For example "মানুষ ভাত খায়**"** (man eats rice). Here "মানুষ" (man) is subject and "খায়" (eat) is verb. Now we check the relationship from VT (Table 3) and find there is a True relation between subject (man) and rice (verb) and indicate OV (Table 4) to S11. As we see ভাত (rice) is member of S11 and hence the sentence is semantically correct. Similarly "গরু ভাত খায়**"** ("Cow eats rice") will be semantically incorrect sentence because OV relation is false. So any sentence displaying relationship which is illogical or irrational will indicate semantically incorrect and therefore

will not be acceptable by the framework that we have proposed. Fig. 1 shows the proposed algorithm for semantic check from Bangla text.

For example "মানুষ ভাত খায়" (man eats rice). Here "মানুষ" (man) is subject and "খায়" (eat) is verb. Now we check the relationship from VT (Table 3) and find there is a True relation between subject (man) and rice (verb) and indicate OV (Table 4) to S11. As we see ভাত (rice) is member of S11 and hence the sentence is semantically correct. Similarly "গরু ভাত খায়" ("Cow eats rice") will be semantically incorrect sentence because OV relation is false. So any sentence displaying relationship which is illogical or irrational will indicate semantically incorrect and therefore will not be acceptable by the framework that we have proposed. Fig. 1 shows the proposed algorithm for semantic check from Bangla text.

III. RELATED WORKS

There have been lots of research on semantic analysis from text of different language[14]-[20 .[But the research of semantic analysis for Bangla language is very few in the literature]12][14][15 .[Soma Paul [14][15] describes an analysis of the unification two verb Bangla sentences (V1 and V2) by using semantic principle of compounding Based on HPSG structure[5][21]; The semantic content of a V2 structure-shares with the content of the V1 that selects the V2 in which the first member (V1) chooses between the conjunctive participial form and the infinitive form and the second member (V2) bears the inflection. Both the member verbs are semantically contentful. [12] present a methodology to extract semantic role labels of Bengali nouns using 5W. The

```
check-semantic(sentence)
begin
//Assumed the table VT[i,j]is created if there is a
//valid relation between subject s_i and verb v_j
 //Split the sentence according to subject as s, verb as v
//and object as o;
 sSet{}:=subject set,  vSet():=verb set and
oSet():=object set;
if (s □ sSet and v □ vSet and o □ oSet) then
  begin
      if (VT[s,v]=T) then return set s
           if (object o □ S) return Correct
            else return Incorrect
      else return Incorrect;
   end
end.
```

Figure 1. Algorithm for semantic check of simple Bangla sentences

5W task seeks to extract the semantic information of nouns in a natural language sentence by distilling it into the answers to the 5W questions: Who, What, When, Where and Why. Beth Levin[16][17 [dicovers the behaviour of a verb ,particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning and thus verb behabiour can be used to probe for linguistically relevant pertinent aspects of verb meanings. Hence Levin classifies over 3,000 English verbs according to shared meaning and behavior into different categories. Massachusetts Institute of Technology published a survey [18] onverb classes and alternations in Bangla, German, English, and Korean language for the purpose of investigating relationship between the semantic and syntactic properties of verbs based on cross-linguistic equivalents of Levin's classes[16]. A modified implemetation of Levin's theory is implemented in[19 [for clustering German Verbs which describes and evaluate the application of a spectral clustering technique to the unsupervised clustering of German verbs] .20 [Detects the semantic errors in Arabic texts using distributed architecture, namely, a Multi Agent System (MAS) .In this paper we have investigated the properties of noun and verb and their relation with valid objects on the basis of universal attributes of different animal, species, and objects like chair, table for simple Bangla Text.

TABLE V. VERB CATEGORIZATION

Sample inputs	Sentences	Results
রহিম স্কুলে যায়। সে সাইকেল চালায়। সাইকেল আকাশে উড়ে। করিম ভাত খায়। সে পড়াশুনা করে। তারা বাসায় থাকে। তারা নিয়মিত নামাজ পড়ে না। রহিম ও করিম ভাই ভাই।	Simple sentence = 06 & Others=02	Error=1 Correct=05 Not detected=02

IV. EXPERIMENTAL RESULTS

Bangla is a complicated language and has a complex structural grammar which we have faced during testing our methodology. As we have developed a methodology to detect the semantic error of simple Bangla text, there are no such contents or corpora which are only in simple format of Bengali grammar. Also it is difficult to find any standard corpus that has some semantic error on the text. For this reason, we have taken different sample testing contents which were built by expert people and individuals to detect the possible semantic errors. Table shows a sample experimental analysis for semantic testing. There are different types of sentences in which some are simple structured of Bangla grammar. Some are different types. As we consider only SOV format, the paragraph contains 6 sentences of this form. 2 sentences of this passage are not matched with this structure, so these sentences cannot be detected by this methodology. By the analysis of this passage, 6 sentences are candidate for semantic error checking and out of these 6 sentences 1 sentence has semantic error ("সাইকেল আকাশে উড়ে।") and rest of the 5 sentences are semantically correct.

V. CONCLUSION

We have presented a methodology to detect the semantic error from simple Bangla sentences. We have categorized the nouns and verbs for Bangla sentences. Although the methodology is for simple sentences of the form SOV but the categorization of nouns and verbs can be used for other forms of Bangla sentences such as complex and compound sentences and even for the multiple verb sentences. It is important and necessary to complete the validation table and object verb relation table for all the verbs and nouns of Bangla language. The performance of the proposed technique greatly depends on this. We believe the proposed algorithm can easily be extended for complex and compound sentences for semantic error detection.

REFERENCES

[1] Bendor E. Sag A. and Wawsow T., "Syntactic Theory: A Formal Introduction", CSLI Publications, Stanford, CA, 1999.

[2] Wechsler. S, "The Semantic Basis of Argument Structure", CSLI Publications, Stanford, CA, 1995.

[3] K. M. Azharul Hasan, Al-Mahmud, Amit Mondal, Amit Saha, "Recognizing Bangla Grammar using Predictive Parser", International Journal of Computer Science & Information Technology , 3(6), pp. 61-73, 2011.

[4] K.M.A Hasan, A.Mondal, A.Saha "A context free grammar and its predictive parser for bangla grammar recognition" 13th International Conference Computer and Information Technology (ICCIT), pp. 87-91, 2010.

[5] Md. Asfaqul Islam, K. M. Azharul Hasan, Md. Mizanur Rahman, "Basic HPSG Structure for Bangla Grammar", Proceedings of the 15th ICCIT, pp. 185-189, 2012.

[6] K. M. Azharul Hasan, Md Sajidul Islam, G. M. Mashrur-E-Elahi, Mohammad Navid Izhar, "Sentiment Recognition from Bangla Text", Technical Challenges and Design Issues in Bangla Language Processing, 2013 .

[7] Das A., and Bandyopadhyay S., "Phrase-level polarity identification for Bengali", International Journal of Computational Linguistics and Applications,1(2),169–181, 2010.

[8] Mohammed Nazrul Islam, Mohammad Ataul Karim, "Bangla Character Recognition Using Optical Joint Transform Correlation", Technical Challenges and Design Issues in Bangla Language Processing, 2013.

[9] Shah Atiqur Rahman, Kazi Shahed Mahmud, Banani Roy, K. M. Azharul Hasan, "English to Bengali Translation Using A New Natural Language Processing (NLP) Algorithm" Proceedings of the ICCIT 2003.

[10] K. M. Azharul Hasan, Muhammad Hozaifa, Sanjoy Dutta, Rafsan Zani Rabbi, "A Framework for Bangla Text to Speech Synthesis", Proceedings of the 16th ICCIT , pp. 60-64, 2013.

[11] Md. Hanif Seddiqui, Muhammad Anwarul Azim, Mohammad Shahidur Rahman, M. Zafar Iqbal, "Algorithmic Approach to Synthesize Voice from Bangla Text", Proceedings of the 5th ICCIT, pp. 233-236, 2002.

[12] Amitava Das, Aniruddha Ghosh and Sivaji Bandyopadhyay, "Semantic role labeling for Bengali using 5Ws", International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE, pp 1-8, 2010.

[13] Frey, T., Gelhausen, M., and Saake, "Categorization of Concerns – A Categorical Program Comprehension Model" In Proceedings of the Workshop on Evaluation and Usability of Programming Languages and Tools at the ACM Onward and SPLASH Conferences, pp. 73-82 2011.

[14] Soma Paul, "Composition of Compound Verbs in Bangla", Proceedings of the workshop on Multi-Verb constructions Trondheim, Summer School, 2003 .

[15] Soma Paul, Dorothee Beermann, and Lars Hellan. "Composition of compound verbs in Bangla." Multi-Verb constructions, 2003.

[16] Beth Levin, "English Verb Classes and Alternations: A preliminary investigation" The University of Shikago Press, Shikago, London, 1993.

[17] Levin, B. and R and Rappaport H. M., "Unaccusativity, at the syntax-lexical semantics interface", Cambridge, Mass.: MIT Press, 1995.

[18] Douglas A., Jones, Robert C., Berwick, Franklin Cho, Zeeshan Khan, Karen T. Kohl, Naoyuki Nomura, Anand Radhakrishnan, Ulrich Sauerland,and Brian Ulicny, " Technical Report on Verb Classes and Alternations in Bangla, German, English, and Korea", Massachusetts Institute of Technology Cambridge, MA, USA, 1993.

[19] Chris Brew and Sabine Schulte Walde "Spectral Clustering for German Verbs", Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, pp. 117-124, 2002.

[20] Chiraz Ben, Othmane Zribi and Mohamed Ben Ahmed, " Detection of semantic errors in Arabic text", Journal of Artificial Intelligence, 195, pp. 249-264, 2013.

[21] A. Copestake, "Implementing Typed Feature Structure Grammars", CSLI Publications, Stanford, 2002.