Extracting Semantic Relatedness For Bangla Words

Abdullah Al Hadi[1], Md. Yasin Ali Khan[2] and Md. Abu Sayed[3]
Department of Computer Science and Engineering,
[1, 2] Chittagong University of Engineering and Technology, Chittagong-4349, Bangladesh
[3] American International University-Bangladesh, Banani, Dhaka-1213, Bangladesh
[1] 10abdullah61@gmail.com, [2] shihabyasin@gmail.com, [3] abusayed93.cse@gmail.com

Abstract— **a framework for extracting semantic relational words in Bangla is presented in this paper. Here extraction of Synonyms, Antonyms, Hyponym, Hypernym, Meronym, Holonym and Polysemy are primarily investigated as a rule based model. For every word two other things: concept and parts of speech category are also presented for clarification. A semantic analyzer is used to extract these relations from nouns, adjectives and verbs.**

Keywords— semantic relatedness,relation extraction, rule based model, semantic similarity.

I. INTRODUCTION

Natural Language Processing (NLP) is a growing field of interest for researchers of computer science, artificial intelligence, linguistics and human computer interaction [1]. Semantic relations are unidirectional underlying connections between concepts because it studies meaning of a language. Language processing consists of morphological, syntactic, semantic and pragmatic analysis steps where semantic relatedness is important. Among two types of semantic approaches 'Compositional Semantics' deals with the meaning of individual units. Then it helps forming larger units. On the other hand 'Lexical Semantics' identify and represent semantics of each lexical item. This helps to understand meaning of larger units. Semantic relatedness has many important applications in inference, reasoning, Question Answering, Information Extraction, Machine Translation and other NLP applications. Actually semantic relations work like building blocks for creating a semantic structure of a sentence. Semantic relatedness implies degree to which words are associated via any relation like synonymy, meronymy, hyponymy, hypernymy, functional, associative and other types of semantic relationships. It has immense application in information retrieval, automatic indexing, word sense disambiguation, automatic text correction etc.

This paper will propose a rule based approach for measuring semantic relatedness between Bangla words. The semantic relatedness between words is computed based on their features they possess using some predefined rules.

II. RELATED WORKS

In literature different works on semantic relatedness and relation extraction are found.

One of the earlier work from Princeton University was WordNet[2] in English by George Miller in 1985. Now it is directed by Christiane Fell Baum[3]. Mentionable other works are FrameNet[4], PropBank[5] and feature based similarity model by[6]. Using multiple information sources semantic similarity between words was investigated by Li et al.[7]. Relations between nominal was investigated by Girju et al.[8] and between noun phrases were investigated by Davidov[9] and Moldovan[10]. Also relation between named entities and clauses were investigated by Hirano et al.[11] and Szpakowicz et al.[12] respectively. Measures of semantic similarity and relatedness in the biomedical domain were investigated by Ted Pedersen et al.[13] Based on corpus statistics and lexical taxonomy similarity was investigated by[14]. Similarity measurement based on web search engine described by Bollegara et al[15] and Cilibrasi et al.[16] for Google. Wikipedia-based semantic relatedness can be found in [17, 18]. Das et al. [19] developed a Semantic Net in Bangla that are basically based on common usage of Bengali people. For Bangla based on Princeton Word Net IIT Bombay gives a miniature idea, only for synonyms. M.Khan proposed some modification there [20].

Getting motivation from above works we would like to propose an automated as well as independent semantic relation extractor with a set of semantic features for Bangla words.

Main investigation of this paper can be stated as:

- To design a semantic relation extractor that can identify the relationships among Bangla words.
- To implement the system by proposing a set of semantic features for Bangla word categories.
- To verify the system for several kinds of Bangla words.

III. SEMANTIC RELATIONS

Theoretically semantic relations can be described by R(x, y) where R is the relation type and x, y are first and second arguments correspondingly.

This section is for discussion about some relations between lexical items.

- **Synonymy**: Refers to words that are pronounced and spelled differently but contain the same meaning. Such as *anondo*(আনন্দ), *ullash*(উল্লাস), *khushi*(খুশি) are synonyms.
- **Antonymy**: Refers to words that are related by having the opposite meanings to each other. Such as *hasi*(হাসি) and *kanna*(কান্না) are antonyms to each other.

978-1-5090-1269-5/16/$31.00 ©2016 IEEE

- **Hyponymy** and **Hypernymy**: Refers to a relationship between a general term and the more specific terms that fall under the category of the general term. For example, the colors *lal(*লাল*), sobuj(*সবুজ*), sada(*সাদা*)* and *holud(*হলুদ*)* are hyponyms. They fall under the general term of *rong(*রঙ*)*, which is the hypernym of the above colors.
- **Polysemy**: A single word or phrase with two or more distinct meanings. For example:
  pata (পাতা): Leaf of tree.
  pata (পাতা): Page of books.
- **Holonymy** and **Meronymy**: A semantic relation that exists between a term denoting whole (the holonym) and a term denoting a part that pertains to the whole (the meronym). For example, *angul*(আঙ্গুল) is a meronym of *hat*(হাত) because *angul*(আঙ্গুল) is part of a *hat*(হাত) and *hat*(হাত) is a holonym of *angul*(আঙ্গুল).

In a language a word may appear in more than one grammatical category and within that grammatical category it can have multiple senses. Lexical semantic relations support the grammatical categories namely Noun (বিশেষ্য), Adjective (বিশেষণ) and Verb (ক্রিয়া).

IV. Mathematical Realization

In this section mathematical description of semantic relatedness will be given.

Let, W1 be the input word and F1 = {f11, f12, f13……..…, f1n} is the set of features of the word W1. Now R is a relation (e.g. synonymy, antonymy, hypernymy, hyponymy, polysemy and holonymy) to find word W2 which should be related to W1 in such a way that W1 and W2 resembles with the definition of R.

$$R\ \{W1\ (F1)\} = W2\ (F2) \quad \text{——————} \quad (1)$$

Meaning W1 and W2 are R related.

Where F2= {f21, f22, f23…, f2n} is the set of features of word W2.

For the relation synonymy, W1 and W2 will share all their features with equal value.

For antonymy, W1 and W2 will share almost all of their features except one and this one contains the reverse value.

For hypernymy, W2 will share almost all of the features of W1 except one and this one defines a general term i.e. it contains a neutral value.

For hyponymy, W2 will share almost all of the features of W1 except one and this one contains a neutral value for W1 and polar (positive or negative) value for W2.

For polysemy, W1 and W2 will be same (i.e. W1=W2) but features are different (i.e. F1 is not exact equal to F2).

For meronymy, W2 will share almost all of the features of W1 except one and this one contains a fractional value in F2.

For holonymy, W2 will share almost all of the features of W1 except one and this one contains a fractional value in F1 but not F2.

V. Methodology And System Architecture

The key objective of our work is to design a semantic relation extractor that can identify different relational words. The schematic representation of our proposed analyzer is illustrated in Fig. 1.

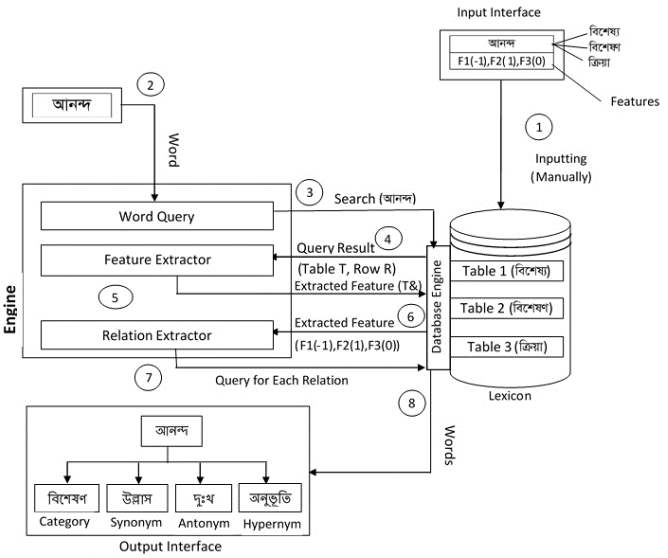

Fig. 1 Schematic representation of proposed system

First of all, some words were selected with effective features to store in the database using input interface. These features are chosen in such a way that they can illustrate how words are both similar and/or different and emphasizes the uniqueness of each word. For example, features for the words "আনন্দ" and "উল্লাস" will be [Animate (-1), Human (+1), Gender (0), Emotion (+1)] and for the word "দুঃখ" [Animate (-1), Human (+1), Gender (0), Emotion (-1)]]. Again, for "অনুভূতি", features will be [Animate (-1), Human (+1), Gender (0), Emotion (0)]. The words may be Noun(বিশেষ্য), Adjective(বিশেষণ) or Verb(ক্রিয়া). In database engine these words will be kept in different tables, since features of each word categories are different. In linguistic, this database engine is called Lexicon which is a dictionary of words where each word contains some syntactic, semantic and some possible pragmatic information.

Example Database tables and their corresponding features are illustrated as below in tables. 2, 3,4,5,6 and 7.

Words / Features	baba	pita	ma	manush	chokh
Countable	1	1	1	1	1
Common	-1	-1	-1	1	0
Animate	1	1	1	1	1.1
Human	1	1	1	1	1.1
Honourable	1	1	1	0	x
Gender	1	1	-1	0	x
Adult	1	1	1	0	x
Material	x	x	x	x	x

Solid	x	x	x	x	x

Table - 2: Example Noun Table from Database

Column Name	Features' Value Description
countable	Countable=1 Uncountable=-1
Common	Common=1 Proper=-1 Neutral = 0 Not applicable = null(x)
Animate	Animate=1 Inanimate=-1 Not applicable = null(x)
Person	Person =1 Neuter = -1 Not applicable = null(x)
Honorable	Honorable=1 Non-honorable=-1 Neutral = 0 Not applicable = null(x)
Gender	Male=1 Female=-1 Neutral = 0 Not applicable = null(x)
Adult	Old/Very Old = 2 Middle Age = 1 Young = -1 Little age / child =-2 Neutral = 0 Not applicable = null(x)
Material	Material = 1 Abstract = -1 Not applicable = null(x)
Solid	Solid = 1 Non-Solid=-1 Not applicable = null(x)

Table - 3: Feature Description of Noun

Words / Features	anondo	ullash	dukkho	valo	chalak	abeg
Animate	-1	-1	-1	1	1	-1
Human	1	1	1	1	1	1
Gender	0	0	0	0	0	0
Quality	x	x	x	1	2	x
Emotion	1	1	-1	x	x	0
Quantity	x	x	x	x	x	x
Size	x	x	x	x	x	x
Beauty	x	x	x	x	x	x

Table - 4: Example Adjective Table from Database

Column Name	Features' value description
Animate	Animate = 1 Inanimate = -1
Human	Human = 1 Neuter = -1
Gender	Male = 1 Female = -1 Neutral = 0
Quality	Good Quality = Positive value(+) Bad Quality = Negative value(-) For distinguishing = 1,2,3,4 Neutral = 0 Not Applicable = Null(x)
Emotion	Good Emotion = Positive value (+) Bad Emotion =Negative value (-) Neutral = 0 Not Applicable = Null(x)
Quantity	Large Quantity = Positive value (+) Small Quantity= Negative value (-) For distinguishing = 1,2,3 Neutral = 0 Not Applicable = Null(x)
Size	Big Size = Positive value (+) Small Size = Negative value (-) For distinguishing = 1,2,3,4 Neutral = 0 Not Applicable = Null(x)
Beauty	Beautiful = Positive value(+) Ugly = Negative value (-) Neutral = 0 Not Applicable = Null(x)

Table - 5: Feature Description of Adjective

Words / Features	Jog_ kora	Biog_ kora	Deoa	Neoa	Prodan_ kora	Poriborton_ kora
Animate	-1	-1	-1	-1	-1	0
Person	1	1	1	1	1	1
Gender	0	0	0	0	0	0
Move	x	x	x	x	x	X
Change	1	-1	2	-2	2	0
State	x	x	x	x	x	x
Decision	x	x	x	x	x	x

Table - 6: Example Verb Table from Database

Column Name	Features' value Description
Animate	Animate = 1 Inanimate = -1
Human	Human = 1 Neuter = -1
Gender	Male = 1 Female = -1 Neutral = 0
Move	In = Positive value(+) Out = Negative value(-) Neutral = 0 Not Applicable = Null(x)
Change	Value upgrading/Possessing/Constructing = Positive value(+) Value degrading/Give up/Destructing = Negative value(-) For distinguishing = 1,2 Neutral = 0 Not Applicable = Null(x)

State	Continuity/Starting = Positive value(+) Discontinuity/Ending = Negative value(-) For distinguishing = 1,2 Neutral = 0 Not Applicable = Null(x)
Decision	Supportive Decision = Positive value(+) Anti-Supportive Decision = Negative value(-) Neutral = 0 Not Applicable = Null(x)

Table -7: Feature Description of Verb

Table 2 to 7 describes in details what features and their corresponding value range had been chosen for Noun, Adjective and Verb correspondingly.

VI. Illustrated Example

Take a sample word, for example "আনন্দ" from user interface (Fig 8) for processing.

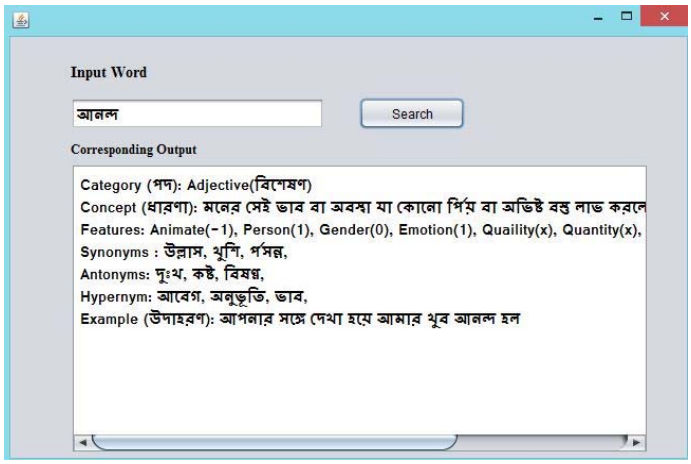

Fig. 8: User Interface

The word will be searched in each table of the Lexicon by Word Query. Queries are the primary mechanism for retrieving information from a database. Many database management systems use the Structured Query Language (SQL) standard query format. Word Query will result a pointer value (Table T and Row R). Feature Extractor will extract all features [Animate (-1), Human (+1), Gender (0), Emotion (+1)] of the pointer (T, R) which are the key element of Relation Analyzer. The analyzer will analyze the extracted feature for each relation. The acceptability of our work is mainly depending on this step. Then the analyzer will build a query from analyzed data to extract the closely related word(s) from Lexicon. For synonym, it will extract the word "উল্লাস" since its features are same [Animate (-1), Human (+1), Gender (0), Emotion (+1)] and for antonym, it will extract the word "দুঃখ" since it's at least one feature [Emotion (-1)] is opposite. Again, it will extract the word "অনুভূতি" for hypernym since a feature [Emotion (0)] is not clearly defined. Then the word(s) and possible some other information (Category, Sub-category, Concept, Example) will be shown in the Output Interface.

VII. Experiments

A. System Requirements

An Intel(R) Core(TM) i3-2100 CPU with 3.10GHz is used having 4GB Ram and 32bit operating system.

B. Implementation:

For designing this system Java is used as computer language and SQLite as Database.

C. Evaluation and Measurement:

For evaluating some words selected randomly and after inputting the words into the system performance had been measured.

D. Limitation:

There is no ideal convention for selecting features. This is totally subjective and dependent highly on application domain.

VIII. Results

For measuring performance of our model engine we choose random sampling method. We randomly selected 80 words for testing, and take a note of number of words where all relations correctly retrieved and number of words where at least one relation incorrectly retrieved. After several experiments we've calculated average number of words where all relations correctly retrieved and average number of words where at least one relation incorrectly retrieved. Then we measure error and accuracy using formulas like below:

$$Error = \{(\text{Average No. of words where all relations correctly retrieved})/80\} * 100;$$

$$Accuracy = 100 - Error;$$

After experimenting randomly with different Bangla words taken from the built in corpora we have seen mentionable performance that are shown in Table. 9.

Word Category	No. of input words(Random Sampling) taken to test	Average No. of words where all relations correctly retrieved	Average No. of words where at least one relation incorrectly retrieved	Error	Accuracy
Noun	80	75	5	6.25%	93.75%
Adjective	80	78	2	2.5%	97.5%
Verb	80	78	2	2.5%	97.5%

Table - 9: Experimental Result

Overall accuracy = 96.25% and Error = 3.75%

The reason for the lower accuracy of nouns is due to its word variation that is not always possible to identify each noun word specifically. Many nouns are very general enough that to identify that noun separately extra one specific feature must be added. By adding more proper features, accuracy of the system may be increased. Major limitation of this work is relatively small size of lexicon compared to other works in non-Bengali languages and also personal influences on selecting features for different types of words as there is no state-of-art rules for it. To best of our knowledge this is the first work in Bengali literature extracting semantic relatedness.

IX. Conclusion

A feature based semantic relatedness system is presented for Bangla. The various semantic features can indicate the semantic structure of a word. It does not depend on specific lexical resources or knowledge representation languages. As it uses own source of data, it maximizes the coverage of possible interpretations. In this work as feature engineering is highly subjective more analytical review may increase performance. Experimental results show satisfactory performance. Future research will be conducted on extending feature set for more lexical units such as noun phrase, multiword expression with more effective features and with more words from Bangla language. It will also be interesting to investigate semantic distances.

References

[1] Wikipedia. “Natural Language processing”, Wikipedia.org. Available:http://en.wikipedia.org/wiki/Natural_langage_processing [Last Modified: 3 March 2015, 16:51].

[2] Miller, George A. WordNet: A Lexical Database for English. Communications of the ACM, 1995, 38:39–41.

[3] C. Leacock and M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet, an Electronic Lexical Database, 1998, pp. 265- 283, MIT Press.

[4] Baker, Collin F., Charles J. Fillmore, and John B.Lowe, “The Berkeley FrameNet Project”. In Proceedings of the 17th International Conference on Computational Linguistics, Montreal, Canada, 1998.

[5] Palmer, Martha, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 2005, 31(1):71–106.

[6] Tversky, A. Features of similarity. Psychological review, 1977, 84(4):327.

[7] Yuhua Li, Zuhair A.Bandar, and David McLean, An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, IEEE Transactions on Knowlege and Data Engineering,vol 15, 2003,pp 871-882.

[8] Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In Proceedings of the Fourth International Workshop on Semantic Evaluations, pp. 13–18, Prague, Czech Republic.

[9] Davidson, Dmitry and Ari Rappaport. Classification of Semantic Relationships between Nominals Using Pattern Clusters. In Proceedings of ACL-08: HLT, 2008, pp. 227–235, Columbus, Ohio.

[10] Moldovan, Dan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. Models for the Semantic Classification of Noun Phrases. In HLT-NAACL 2004: Workshop on Computational Lexical Semantics, 2004, pp. 60–67.

[11] Hirano, Toru, Yoshihiro Matsuo, and Genichiro Kikui. Detecting Semantic Relations between Named Entities in Text Using Contextual Features. In Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions, 2007, pp. 157–160.

[12] Szpakowicz, Barker, Ken Barker, and Stan Szpakowicz. Interactive semantic analysis of ClauseLevel Relationships. In Proceedings of the Second Conference of the Pacific ACL,1995, pp. 22–30.

[13] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan and Christopher G.Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics,vol 40, 2007, pp. 288-299.

[14] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th international conference on research in computational linguistics, 1997,pp. 19–33, Taipei, Taiwan.

[15] Danushka Bollegara, Yutaka Matsuo, and Mitsuru Isizuka, Measuring Semantic Similarity between Words Using Web Search Engines, Proceedings of the 16th International World Wide Web Conference (WWW2007),pp. 757-766, Banff, Alberta, Canada, 2007.

[16] Rudi L. Cilibrasi and Paul M.B. Vitanyi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, vol 19, 2007, pp. 370-383.

[17] Evgeniy Gabrilovich and Shaul Markovitch, Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), 1606-1611, Hyderabad, India, 2007.

[18] Michael Strube and Simone Paolo Ponzetto, WikiRelate! Computing Semantic Related-ness Using Wikipedia, Proceedings of the 21st National Conference on Artificial Intelligence, 2006, pp.1419-1424, Boston, Mass.

[19] Das, A. and Bandyopadhyay, S. (2010). Semanticnet-perception of human pragmatics. In Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon, pp. 2–11, Beijing, China. Coling 2010 Organizing Committee.

[20] Kamrul Hayder, Naira Khan, and Mumit Khan, "Bangla WordNet Development Challenges and Solutions", Center for Research on Bangla Language Processing, October 8, 2007, BRAC University.