XBLQPS: An Extended Bengali Language Query Processing System for e-Healthcare Domain

Kailash Pati Mandal[1], Prasenjit Mukherjee[2], Atanu Chattopadhyay[3], Baisakhi Chakraborty[4]
Computer Science and Engineering, National Institute of Technology, Durgapur, India[1, 2, 4]
BBA (H) and BCA (H) Department, Deshabandhu Mahavidyalaya, Chittaranjan, India[3]

***Abstract*—The digital India program encourages Indian citizens to become conversant with e-services which are primarily English language-based services. However, the vast majority of the Indian population is comfortable with vernacular languages like Bengali, Assamese, Hindi, etc. The rural villagers are not able to interact with the Relational Database Management system in their native language. Therefore, create a system that produces SQL queries from natural language queries in Bengali, containing ambiguous words. This paper proposes a Bengali Query Processor named Extended Bengali language Query Processing System (XBLQPS) to handle queries containing ambiguous words posted to a Healthcare Information database in the electronic domain. The Healthcare Information database contains doctor, hospital and department details in the Bengali language. The proposed system provides support for the Bengali-speaking Indian rural population to efficiently fetch required information from the database. The proposed system extracts the Bengali root word by removing the inflectional part and categorizing them to a specific part of speech (POS) using modified Bengali WordNet. The proposed system uses manually annotated parts of speech detection of a word based on Bengali WordNet. Patterns of noun phrases are generated to detect the correct noun phrase as well as entity and attribute(s). Entity and attributes are used to prepare the semantic table which is utilized to create the Structured Query Language (SQL). The simplified LESK method is utilized to resolve ambiguous Bengali phrases in this query processing system. The accuracy, precision, recall and F1 score of the system is measured as 70%, 74%, 73%, and 73% respectively.**

Keywords*—Relational database management system (RDBMS); modified bengali WordNet; LESK algorithm; structured query language (SQL); natural language query

I. INTRODUCTION

Indian citizens have become more familiar and conversant with online or electronic systems like e-banking, e-governance, e-health, e-tourism, and e-education due to the empowerment through the digital India program. Nowadays most government facilities are electronic based. However, the vast majority of the Indian population is village based. There are 0.7 million villages in India. Most villagers are not accustomed to the English-based systems. They are comfortable with their own vernacular language(s) and Bengali is one of them. Bengali is widely used in West Bengal, Andaman and Nicobar Islands, Tripura, Assam, and other states. The Bengali language is the official language of Bangladesh, West Bengal and Tripura. One of the major e-service requirements in the rural interiors is related to healthcare systems. Health care domain-based e-services in vernacular languages are a really challenging task. Today, the query-response model in vernacular language is an open research problem in the research community. This is because the resources of vernacular languages are very low and difficult to implement. A good query response model in the Bengali language can be helpful for naive users to extract information without any technical knowledge. WordNet plays a vital role to develop the query-response model. The Query expansion (QE) is a well-known technique used to enhance the effectiveness of information retrieval. A new approach for QE using Wikipedia and WordNet as data sources is described in [1]. The IndoWordNet is an important lexicographic resource for different Indian languages. The lexical matrix has been used to construct the relational semantics as discussed in [2]. In the article [3], The BWN is a WordNet database for the Bengali language. It consists of lexical source files, a grinder, a WordNet database, and an interface. Using the WNDB interface, users can interact with the BWN documents in various ways as explained in [3]. The African Wordnet Project aims to develop aligned wordnets for African languages spoken in South Africa. The focus of this article will be on isiZulu as one of the selected languages used in [4]. The meaning of an ambiguous word has been determined according to its context. The proposed method generates context by comparing a doubtful word with words in the input document for similarity. The similarity computation is based on BabelNet's semantic framework. [5].

The naive users who access the online Healthcare Information Management System in the Bengali language is a challenging task. Therefore, this paper proposes a modified Bengali WordNet and a natural language query-response system, namely Extended Bengali language Query Processing System (XBLQPS) that can handle natural language queries in Bengali. This automated system generates responses from the knowledge database. A relational Database Management System (RDBMS) is used to implement the knowledge database. This work is an enhancement of the work of [6], where a Query Processing System for the Bengali language has been proposed and discussed.

This research article consists of various sections. Section II narrates related works. Section III describes the objective of the proposed system where Section IV explains the proposed system and Section V describes analysis and discussion of the proposed system. Section VI is about the modified Bengali WordNet and health information database. Section VII illustrates the used tool and IPA notation. Section VIII derives the time complexity of the proposed system. Section IX is related to analyze the precision, recall, and accuracy of the

XBLQPS, Section X presents the limitations and future works. The proposed system is concluded in Section XI.

II. Related Work

Bengali WordNet databases are useful and a few good implementations have been done by Pushpak Bhattacharyya [2] and Farhana Farque [3]. The proposed work is influenced from Pushpak Bhattacharyya [2], Farhana Farque [3] and A. Haque [22].

The article [7] stated a question-answering model using natural language query where interactions between table and question are complex type. The sketch-based approach has been applied to solve the complexity as in [7]. Word sense disambiguation (WSD) is to find the proper meaning of a word in any context. It involves incorporating word knowledge from external knowledge resources to remove equivocalness. A WSD tool has been implemented using Hindi WordNet where it is a knowledge-based approach as in [8]. In the paper [9], the sense Induction approach has been used in the algorithm for word sense disambiguation (WSD) in Bengali. Ten frequently used Bengali ambiguous terms and 200 phrases of each term are utilized to test the WSD model. The radix tree-based structure is used to keep context information as well as faster searching of a word in the paragraph. The WordNet is used as a knowledge base to disambiguate the word as discussed in [10]. The Case-Based Reasoning interpretation technique was used to decipher the confusing word Gurmukhi, also known as Punjabi, in Indian Regional Language. The solution to the new problem was inferred from the previously solved problems as described in [11]. Machine translation at the human level can be helpful greatly using WSD. Through the FP-Growth method, Authors [12] provide a system for WSD in their study as in [12]. The fuzziness of semantic relation has been applied in Fuzzy Hindi WordNet (FHWN). Various membership values of semantic relations of the FHWN were considered to extract the correct sense as in [13]. In the article [14], the authors described the root word extraction technique from the Bengali inflected nouns by applying well-defined grammatical mapping rules (GMRs) between nominal bases and inflections. The proposed methodology can be applied to the Bengali grammar and inflected words that were described in [14]. To decipher the confusing term, the LESK algorithm has been utilized. The result set of the proposed system is in line with the result set of KBBI as well as provides an accuracy of 78.6% for one of the ambiguous words while 62.5 % for two ambiguous words as explained in [15]. The effort of identifying the meaning of a word in a certain situation is known as word sense disambiguation (WSD). The novel WSD [16] model has been introduced where the proposed model calculates each word's meaning uniquely. The creation of Arabic WordNet (AWN) has made lexical resources available to the Arabic NLP community. The usage of this resource cannot be considered because there are fewer AWN Synsets than other WordNets that has been elaborated in [17]. In the article [18], Vietnamese and Korean both are morphologically rich languages. The high homograph rate in the Korean language is word ambiguities that affect neural MT (NMT). There isn't a sufficient, publicly accessible parallel Korean and Vietnamese that can be utilized to train translation models as implemented in [18]. A hybrid approach is used for Urdu word stemming that is helpful for information extraction, textual categorization, data analysis and related applications. This proposed approach [19] works on unigram, bigram, and trigram features that were discussed in [19]. The word was disambiguated by combining the sense relativeness algorithms with a neural model in the paper [20]. The proposed model works on POS-labeled text corpus and the length of the context may be varied as discussed in [20]. The raw collection of Bangla text has been used to generate meaning-tagged data. The Bangla meaning tagged data contains root word form and their POS type of an ambiguous word with 86.95% performance as implemented in [21]. The Bangla Word Sense Disambiguation System can be distinguished by some confusing Bangla phrases. Parsing and detection are two main working phases of the proposed system [22]. An Algorithm has been developed that clears the confusing word according to the categories of ambiguous words that are nouns, adjectives, and verbs as in [22]. Authors were focused on how to differentiate the important records and generate a summary from them. In the proposed system, the authors applied natural language processing, WordNet and lexical chains for the summary generation of a text as explained in [25]. An Arabic WordNet (AWN) has been used to overcome the WSD problem where word semantic similarity has been checked by multiple Arabic stemming algorithms. This work is related to reducing the gap in Arabic NLP compared to English as in [26]. Table I shows the results of a comparative investigation of existing systems with the proposed system.

TABLE I. A Comparison of Identical Systems using the XBLQPS

SL NO.	AUTHOR(S) & SYSTEM	METHODOLOGY USED IN SAME TYPE SYSTEM	METHODOLOGY APPLIED IN THE PROPOSED SYSTEM (XBLQPS)
1	S. Basuki et al. & LESK Algorithm Utilization for Word Sense Disambiguation (WSD) for Indonesian Homograph Word Meaning Determination [15]	1) The LESK procedure has been used to disambiguate Indonesian Homograph Word referred in [15]. 2) This system provides 78.6% accuracy if one ambiguous word present, and 62.5% accuracy if two ambiguous words present. 3) The time complexity has not been discussed for this system	1) The proposed system uses the LESK algorithm, modified Bengali WordNet and formation of patterns to disambiguate the Bengali word. 2) The overall accuracy of XBLQPS is 70%. 3) The time complexity has been calculated for the proposed system.
2	M. S. Kaysar et al.& Applying FP-Growth Algorithm to Disambiguate Bengali Words [12]	1) The FP-Growth Algorithm has been used to disambiguate the Bengali ambiguous word. 2) This system provides 80% accuracy. 3) The time complexity has not been specified here.	1) The LESK has been used to disambiguate the Bengali ambiguous word. 2) The XBLQPS provides 70% accuracy. 3) The complexity has been computed here.

3	D. O et al. & Word Sense Disambiguation Utilizing Word Vector Representation from a Knowledge-based Graph Based on Word Similarity Calculation [5]	1) The system referred in [5] has been used to disambiguate the ambiguous word using similarity calculation with help of semantic network structure of BabelNet. 2) This system does not generate pattern(s). 3) The precision, recall and F1 score on semEval-2013 dataset are 75%, 75%, and 75% respectively whereas on semEval-2015 are 69.2%, 62.6% and 65.8%.	1) The XBLQPS uses LESK algorithm and modified Bengali WordNet to disambiguate the ambiguous word. 2) The XBLQPS generates pattern(s). 3) The precision, recall and F1 score for the proposed system are 74%, 73% and 73% respectively.
4	M. Biswas et al. & Construction of a Bangla Sense Annotated Corpus for Disambiguation of Word Sense [21]	1) The system referred in [21] creates sense annotated corpus containing ambiguous word. 2) The accuracy of sense annotated corpus is 86.95%. 3) The time complexity has not been discussed for this system.	1) The XBLQPS creates SQL from the Bengali language query containing ambiguous word. 2) The proposed system provides 70% accuracy. 3) The time complexity of XBLQPS has been computed.
5	K. P. Mandal et al. & An unique Bengali Language Query Execution System in the field of health [6]	1) The system referred in [6] does not able to handle the query containing ambiguous word. 2) This system does not use LESK algorithm. 3) The time complexity of this system is $O(n^4)$.	1) The XBLQPS can process the query containing ambiguous word. 2) This system uses LESK algorithm. 3) The time complexity of the proposed system is $O(n^3)$.
6	C. Lachichi et al. & Machine translation and external linguistic elements have been used to enrich Arabic WordNet[17]	1) The system referred in [17] uses Machine Translation and External Linguistic Resources to enhance the Arabic WordNet. 2) It is an Arabic WordNet enhancement system. 3) The average accuracy of this system is 0.48.	1) The XBLQPS uses Bengali WordNet to disambiguate the Bengali ambiguous word. 2) It is a Bengali language query processing system. 3) The accuracy of the XBLQPS is 0.70.
7	A. Haque et al. & Utilizing dictionary-based approach, a Bangla word sense disambiguation model has been proposed [22]	1) The system referred in [22] is a dictionary based Bengali word sense disambiguation system. 2) The accuracy of this system is 82.40%. 3) This system does not convert Bengali natural query to SQL.	1) The XBLQPS is a query processing system containing Bengali ambiguous word. 2) The accuracy of the XBLQPS is 70%. 3) The XBLQPS converts Bengali natural language query to SQL.

III. Objective of the Proposed System

- The main objective of this research is to create a system that can handle natural language queries in Bengali in the healthcare domain. A naïve user will be able to extract healthcare information without any technical knowledge.
- The proposed system will be able to handle natural language queries that are containing ambiguous words in Bengali.
- The proposed system will be able to generate a response from the Bengali Query that contains ambiguous words.
- The proposed system will generate SQL from natural language queries in Bengali without any manual intervention.

IV. Proposed XBLQPS

Following are the steps of the proposed system:

Step 1: The user logs into the proposed system and submits the Bengali language query.

Step 2: The system slices the query into token(s).

Step 3: The root word extraction and POS tagging are done with the help of modified Bengali WordNet.

Step 4: Once the POS tagging is over, the proposed system generates pattern(s) with the help of noun and noun phrases.

Step 5: The ambiguous term is disambiguated using the simplified LESK method. The proper sense of an ambiguous Bengali word was discovered by examining the most significant number(s) of common words existent between the word's current context and gloss.

Step 6: After identification of the correct sense of every pattern, entity and attribute are also determined from the table named “table_entity_attribute_sensing” for semantic analysis.

Step 7: Once entity and attribute are detected, a set of predefined rules are used to generate an SQL query.

Step 8: After generating the SQL query, the proposed system executes and retrieves desired information from the health information database.

There are three tables in the health information database. "table_hospital" contains the hospital's name and address. The "table_doctor" includes the doctor's name, qualification, specialization, fee, and a particular doctor connected with the department and hospital. The “table_department” contains the department’s name and which department exists in which hospital.

Finally, the desired result sends to the user. The workflow diagram of the XBLQPS is depicted in Fig. 1. The workflow diagram (Fig. 1) shows the functionality of each component of the proposed system.

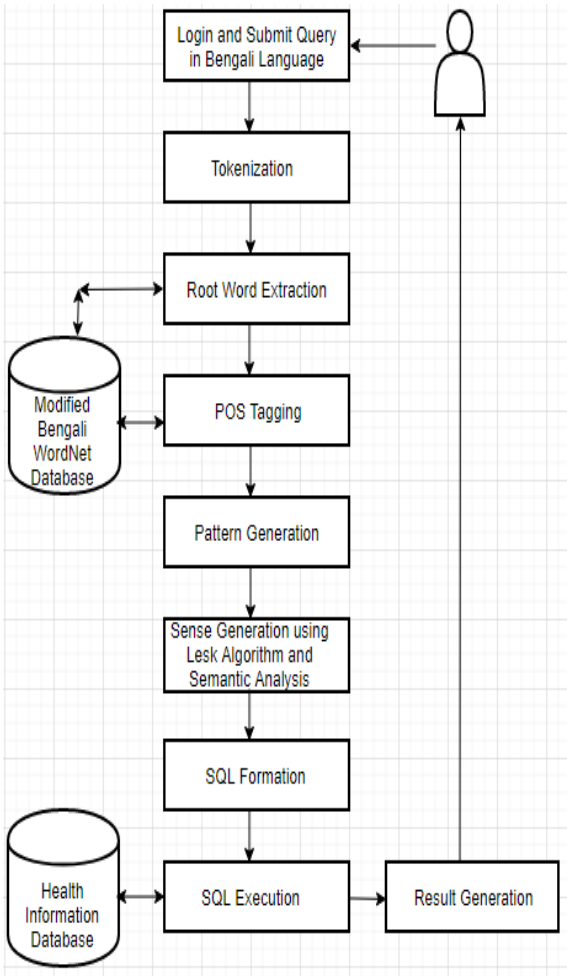

Fig. 1. Work Flow Diagram of XBLQPS.

V. Analysis and Discussion of the Proposed System

The detailed analysis and description of each component of the proposed system are as follows.

Component 1: Login into the system and submit the query in the Bengali language.

The user logs in to the proposed system and submits the query string in the Bengali language. For instance, a query has been provided. The example of query 1 is given in Fig. 2.

আসানসোলে মাথা খারাপের চিকিৎসার জন্য হাসপাতাল কোথায় আছে?

(asansole mat^{h}a k^{h}araper tʃikitsar dʒno ɦaspatal kot^{h}aę atʃ^{h}e?)

(Where is a hospital for treatment of mental diseases in Asansol?)

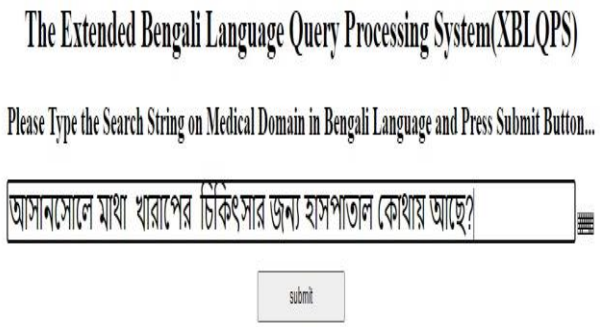

Fig. 2. Example Query 1 in Bengali.

Component 2: Tokenization

The query string will be sliced into small linguistic units called tokens after removing punctuation mark(s). These token(s) are stored in a string array. Table II shows the array indexing in the string array of the user-submitted query after tokenization.

Component 3: Root word extraction

The nominal word is the main backbone of SQL formation. Often the noun exists in the inflected form in the Bengali sentences as Bengali is a highly inflected language. Before POS tagging and extraction of correct sense, each token is compared with the inflectional part table i.e. "table_inflectional_part" which is shown in Table XV in the modified Bengali WordNet DB. The proposed system compares each token with the inflectional part table, that is, Table XV. If any of the inflectional parts in the inflectional part table matches with the trailing part of the token, then the XBLQPS removes the matched part from the token. In this way, the proposed system extracts the root word by removing the matched part from the token. If the trailing part of the token does not match with the inflectional part table in the Modified Bengali WordNet DB, then the XBLQPS will consider that the token itself will be a root word. After removing the inflectional part, all token(s) will be stored again in another string array. After extraction of the root word, the string array has been given in Table III.

TABLE II. Tokenization

array index	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
String array of user query	আসানসোলে (asansole)	মাথা (mat^{h}a)	খারাপের (k^{h}araper)	চিকিৎসার (tʃikitsar)	জন্য (dʒno)	হাসপাতাল (ɦaspatal)	কোথায় (kot^{h}aę)	আছে (atʃ^{h}e)

TABLE III. Root Word Extraction

array index	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
String array of root word	আসানসোল (asansol)	মাথা (mat^{h}a)	খারাপ (k^{h}arap)	চিকিৎসা (tʃikitsa)	জন্য (dʒno)	হাসপাতাল (ɦaspatal)	কোথায় (kot^{h}aę)	আছে (atʃ^{h}e)

Component 4: POS tagging

The proposed system utilizes a POS string array (Table IV) with the same length as the string array of the root words (Table III) for keeping track of parts of speech of every token. The POS string array (Table IV) is helpful for pattern generation of noun phrases. The proposed system selects each token and compares it with the table named "table_word" (Table IX) in the modified Bengali WordNet DB. If the token matches with the table named "table_word" in the modified Bengali WordNet DB, then the corresponding "word_id" field value will be retrieved. The retrieved "word_id" will be compared with the "word_id" field of the table named "table_sense" is shown in Table XI, from where the proposed system will map the "synset_id". Using "synset_id", the XBLQPS will retrieve the "parts_of_speech" field value from the table "table_synset"is given in Table X and that will be inserted into the POS string array (Table IV). If the token does not match with the modified Bengali WordNet DB, then the token will be treated as noun (বিশেষ্য) and that will be inserted into the POS string array (Table IV) as a noun (বিশেষ্য). The selected token's array index value is same in POS string array (Table IV) and string array of root word (Table III). Each token of the query will be compared in the modified Bengali WordNet DB and on finding a match with a word, it will be stored in POS string array (Table IV) in the form of its parts of speech derived from the parts of speech definition given alongside the word in the modified Bengali WordNet proposed in the paper, except the token আসানসোল (Asansol) which is a proper noun; the name of a place. That why the POS string array (Table IV) value of আসানসোল is treated as a noun (বিশেষ্য), which means all unknown tokens which do not match with the modified Bengali WordNet DB will be treated as a noun (বিশেষ্য). The string array after pos tagging has been given in Table IV.

Component 5: Pattern generation of noun phrases

The noun and noun phrases are very much important for entity and attribute(s) identification which are main components of SQL formation. The correct noun or noun phrases have been identified by pattern generation because often the Bengali noun phrase is consisting of more than one consecutive noun (বিশেষ্য) or noun with an adjective (বিশেষণ). The proposed system will generate pattern(s) for the consecutive noun (বিশেষ্য) or consecutive any combination of a noun (বিশেষ্য) and adjective (বিশেষণ). More than one consecutive noun (বিশেষ্য) or combination of noun (বিশেষ্য) and adjective (বিশেষণ) will generate more than one pattern whereas nonconsecutive noun (বিশেষ্য) will generate single pattern. The token which is not belong the noun (বিশেষ্য) or adjective (বিশেষণ) is simply ignores by the proposed system. The system will generate more than one pattern for array index 0, 1, 2 and 3 of the POS string array (Table IV) because this array index contains combination of noun (বিশেষ্য) and adjective (বিশেষণ). The single pattern will generate for array index 5 because this is a non-consecutive noun (বিশেষ্য). During pattern generation, the token order occurrence will be maintained by the proposed system and that will be stored in a string array. The string array of pattern has been given in Table V. The generated patterns for the user given query are as follows.

1) আসানসোল (asansol)
2) মাথা (mat^{h}a)
3) খারাপ (k^{h}arap)
4) চিকিৎসা (ʧikitsa)
5) আসানসোল মাথা (asansol mat^{h}a)
6) মাথা খারাপ (mat^{h}a k^{h}arap)
7) খারাপ চিকিৎসা (k^{h}arap ʧikitsa)
8) আসানসোল মাথা খারাপ (asansol mat^{h}a k^{h}arap)
9) মাথা খারাপ চিকিৎসা (mat^{h}a k^{h}arap ʧikitsa)
*10)*আসানসোল মাথা খারাপ চিকিৎসা (asansol mat^{h}a k^{h}arap ʧikitsa)
*11)*হাসপাতাল (ɦaspatal)

Component 6: Sense extraction of pattern(s) using LESK algorithm and semantic analysis

After generation of pattern(s), every pattern will be compared with the value of "word_name" field of the table named "table_word" (Table IX) in the modified Bengali WordNet DB. If the pattern matches with the value of "word_name" field of the table named "table_word", then corresponding "word_id" will be fetched. The fetched "word_id" will be compared with the value of "word_id" field of the table named "table_sense" is shown in Table XI. If the match is found then corresponding synset_id will be retrieved. Finally, the retrieved "synset_id" will be compared with the value of "synset_id" of the table named "table_synset" is indicated in Table X. If the match is found then the corresponding "parts_of_speech", "gloss_concept_defination", "example_sentence", "meaning" and "possible_attribute" field's values will be retrieved from the table named "table_synset". The retrieved "possible_attribute" value will be compared with the value of "attribute_name" field of the table named "table_entity_attribute_sensing" is specified in Table XIV. If the"possible_attribute" field's value of the table named "table_synset" matches with any value of "attribute_name" field of table named "table_entity_attribute_sensing", then corresponding row value will be fetched from "table_entity_attribute_sensing".

TABLE IV. POS TAGGING

Array Index	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
POS string array	বিশেষ্য (biʃeʃɔ)	বিশেষ্য (biʃeʃɔ)	বিশেষণ (biʃeʃn)	বিশেষ্য (biʃeʃɔ)	অব্যয় (oboe̯)	বিশেষ্য (biʃeʃɔ)	সর্বনাম (sr[2]bonam)	ক্রিয়া (kir[2]ia)

TABLE V. PATTERN GENERATION OF NOUN PHRASES

Array Index	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
String array of patter	আসান সোল (asans ol)	মাথা (mat^{h}a)	খারাপ (k^{h}arap)	চিকিৎসা (ʧikitsa)	আসানসোল মাথা (asansol mat^{h}a)	মাথা খারাপ (mat^{h}a k^{h}arap)	খারাপচিকিৎসা (k^{h}arap ʧikitsa)	আসানসোলমাথা খারাপ (asansol mat^{h}ak^{h}arap)	মাথা খারাপচিকিৎসা (mat^{h}a k^{h}arap ʧikitsa)	আসানসোলমাথা খারাপ চিকিৎসা (asansol mat^{h}a k^{h}arap ʧikitsa)	হাসপাতাল (ɦaspatal)

The row which will be fetched from "table_entity_attribute_sensing" that will be inserted into "table_semantic"is given in Table XVI for semantic analysis where the value of "meaning" field of the tablenamed "table_synset" will be inserted into the "value" field of the table named "table_semantic", If anytime exists NULL value in the "meaning" field then the "value" field of the table named "table_semantic" will also be NULL. Sometimes the pattern may contain more than one value at the "possible_attribue" field. The proposed system fetches all values of "possible_attribute" field and compared with value of "attribute_name" field of the table named "table_entity_attribute_sensing". The matched row values are fetched from "table_entity_attribute_sensing", inserted into "table_semantic" for semantic analysis. If the pattern contains NULL value at their "possible_attribue" field then the pattern simply ignores by the proposed system. If a "word_id" contains only one "synset_id" in the table named "table_sense", that means the pattern is unambiguous .If a"word_id" contains more than one "synset_id" in the table named "table_sense", that means the pattern is ambiguous. To identify the correct sense of the ambiguous pattern, the XBLQPS considers user given query as well as value of the "gloss_concept_defination" and "example_sentence" fields of the table named "table_synset". Next step, the system applies LESK algorithm referred in [27] to find out correct sense of ambiguous pattern(s). The XBLQPS counts the number of common word(s) by comparing between current user given query with the value of "gloss_concept_defination" field of the table named "table_synset" as well as the value of "example_sentence field". The proposed system adds up those above-mentioned count values and selects the row value from the table named "table_synset" which gives maximum count value. The pseudo code of the simplified LESK algorithm is shown in Table VI.

TABLE VI. THE PSEUDO CODE OF THE SIMPLIFIED LESK ALGORITHM

```
Function LESK_simple(word, sentence)
Ideal sense=widely used sense of a word
Maximum common word=0
Context=number of word present in the sentence
Do select every meaning from a set of meaning of every word
Favourite=a set of words present in the gloss example
Collision=compute_collision(favourite, context)
If collision > maximum common word then
Maximum common word=collision
Ideal=meaning
End [27].
```

Eleven patterns have been generated from the user given query; among them five patterns i.eমাথা (mat^{h}a), খারাপ (k^{h}arap), চিকিৎসা (ʧikitsa), মাথা খারাপ (mat^{h}a k^{h}arap) and হাসপাতাল (ɦaspatal) match with table named "table_word". The remaining six patterns i.e. আসানসোল (asansol), আসানসোল মাথা (asansol mat^{h}a), খারাপ চিকিৎসা(k^{h}arap ʧikitsa), আসানসোল মাথা খারাপ (asansol mat^{h}a k^{h}arap), মাথা খারাপ চিকিৎসা (mat^{h}a k^{h}arap ʧikitsa) and আসানসোল মাথা খারাপ চিকিৎসা (asansol mat^{h}a k^{h}arap ʧikitsa) do not match with table named "table_word" and the proposed system treats these mismatched pattern(s) as a value. These values will be compared with the three tables named "table_hospital", " table_department" and "table_doctor" are shown in Table XVII, Table XVIII and Table XIX respectively. These values will be compared with the values present in these three tables and on finding a match with any value of among these above mention three tables, the corresponding attribute name will be retrieved. The retrieved attribute name will be again compared with value of "attribute_name" field of the table named "table_entity_attribute_sensing". If the retrieved attribute name matches with any value of "attribute_name" field of table named "table_entity_attribute_sensing", that corresponding row value will be fetched from "table_entity_attribute_sensing" and inserted into "table_semantic" for semantic analysis. If these values do not match with above mentioned three tables, they are simply ignored by the proposed system.

Among these six patterns, only the pattern আসানসোল (asansol) matches with the value of "hos_add" field of the table named "table_hospital", the attribute name i.e. "hos_add" will be retrieved. The retrieved "hos_add" attribute again will be compared with the value of "attribute_name" field of the table named "table_entity_attribute_sensing". If the "hos_add" matches with "attribute_name" field of the table named "table_entity_attribute_sensing", then the row value will be fetched and will be inserted into "table_semantic" for semantic analysis. Those patterns which match with table named "table_word", some of them are ambiguous and some are unambiguous.

Patterns খারাপ (k^{h}arap), চিকিৎসা (ʧikitsa) and হাসপাতাল (ɦaspatal) have single "synset_id" means these patterns are unambiguous and their corresponding "possible_attribute" field's value will be retrieved. These retrieved "possible_attribute" will be compared with value of "attribute_name" field of the table named "table_entity_attribute_sensing". If the retrieved attribute name matches with any value of "attribute_name" field of table named "table_entity_attribute_sensing", then corresponding row value will be fetched from "table_entity_attribute_sensing". The row which is fetched from "table_entity_attribute_sensing", will be inserted into "table_semantic" for semantic analysis. Patterns খারাপ (k^{h}arap) and চিকিৎসা (ʧikitsa) contain NULL value at their

"possible_attribute" field that why these patterns will be ignored by the proposed system. Pattern হাসপাতাল (haspatal) contains "hos_name" at their "possible_attribute" field. The value "hos_name" is compared with "attribute_nane" field value of the table named "table_entity_attribute_sensing", then the row value is fetched and inserted into the "table_semantic" for semantic analysis. But patterns মাথা (mathˢa) and মাথা খারাপ (mathˢa khˢarap) contains more than one synset_ids means these patterns are ambiguous. Now the proposed system applies the LESK algorithm on patterns মাথা (mathˢa) and মাথা খারাপ (mathˢa khˢarap) to extract the correct sense.

The pattern মাথা(mathˢa) contains more than one sense in the modified Bengali WordNet DB. Here the proposed system uses retrieved "gloss_concept_defination" and "example_sentence" field value from the table named "table_synset" that has been given below.

Sense 1: gloss_concept_defination –কোন পশুর শরীরের উপরের অংশ যেখানে চোখ,নাক, কানযুক্ত থাকে (the upper part of body which contains eyes, nose, ears etc.)

example_sentence –মাথার চিকিৎসা মনোরোগবিদ্যা বিভাগ আছে সেই হাসপাতালে করানো ভালো (It is better to treat mental diseases in a hospital which has psychiatric department.)

Sense 2: gloss_concept_defination –কোন পশু যে দল, সংস্থা অথবা দেশ নিন্ত্রয়ণ করে

matha – head (head of family, head of a state, leader of a nation, head of a group of humans or animals who controls others.).

example_sentence –গ্রামের মাথাদের কথাই শেষ কথা (The words of the head of a village is final). Here head of a village is human being.

The user given query will be compared with the "gloss_concept_defination" as well as "example_sentence" field value of the pattern মাথা. The sense 1 contains maximum number of overlapping word(s). These overlapping words are মাথা, চিকিৎসা and হাসপাতাল. But the sense 2 does contain only one overlapping word i.e. মাথা. So the sense 1 will be closest sense of the pattern মাথা and the value of "possible_attribute" field of that row of the "table_synset" is to be selected. The retrieved possible attribute name is again compared with value of "attribute_name" field of the table named "table_entity_attribute_sensing". If the retrieved attribute name matches with any value of "attribute_name" field of the table named "table_entity_attribute_sensing", then corresponding row value will be fetched from the "table_entity_attribute_sensing". The row which will be fetched from "table_entity_attribute_sensing", and that will be inserted into "table_semantic" for semantic analysis.

The pattern মাথা খারাপ (mentally challenged) contains more than one sense in the modified Bengali WordNet DB. Similarly, the proposed system will retrieve "gloss_concept_defination" and "example_sentence" field value from the table named "table_synset" that has been given below.

Sense 1: gloss_concept_defination –কোন জীব যে উন্মাদ বা মানসিক অসুস্থ (a mentally challenged living being.)

example_sentence –মাথা খারাপের রোগীকে মানসিকরোগের হাসপাতালের ডাক্তার দ্বারা চিকিৎসা করানো দরকার (mentally challenged people should be treated by psychiatric doctor)

Sense 2: gloss_concept_defination –কোন অনিশ্চতার সমন্ধে উদ্বেগ (to be unusually anxious over a trivial matter.)

example_sentence –তুচ্ছ ব্যপার নিয়ে তুমি মাথা খারাপ করোনা (Do not be anxious over such a small matter.)

The user given query will be compared with "gloss_concept_defination" as well as "example_sentence" field value of the pattern মাথা খারাপ. The sense 1 contains maximum number of overlapping word(s). The overlapping word is চিকিৎসা. But the sense 2 does not contain any overlapping word(s). So the sense 1 will be closest sense of the pattern মাথা খারাপ and the value of "possible_attribute" field of that row of the "table_synset" is to be selected. The retrieved possible attribute name will be compared again with value of "attribute_name" field of the table named "table_entity_attribute_sensing". If the retrieved attribute name matches with any value of "attribute_name" field of table named "table_entity_attribute_sensing", then corresponding row value will be fetched from "table_entity_attribute_sensing". The row which will be fetched from "table_entity_attribute_sensing", and that will be inserted into "table_semantic" for semantic analysis. Our proposed system has disambiguated two Bengali ambiguous words মাথা and মাথাখারাপ asমনোরোগ (mental disease) for current context. Therefore two same entries occur for id 4 and 6 in Table VII. The instance of "table_semantic" for above mention user query has been given in Table VII.

TABLE VII. SENSE EXTRACTION OF PATTERN (S) USING LESK ALGORITHM AND SEMANTIC ANALYSIS

id	entity	attribute_name	primary_key	foreign_key	candidate_key	value
1	table_hospital	hos_add	hos_id			আসানসোল(asansol)(propoer Noun)
2	table_hospital	hos_name	hos_id			
3	table_department	dept_name	dept_id	hos_id		মনোরোগ(mnor[2]og)(mental disease)
4	table_doctor	doc_specialist	doc_id	hos_id	dept_id	মনোরোগ(mnor[2]og)(mental disease)
5	table_department	dept_name	dept_id	hos_id		মনোরোগ(mnor[2]og)(mental disease)
6	table_doctor	doc_specialist	doc_id	hos_id	dept_id	মনোরোগ(mnor[2]og)(mental disease)

Component 7: SQL formation

The proposed system will generate SQL from "table_semantic". The proposed system will consider only one entry if more than one rows have the same value in entity, "attribute_name", "primary_key", "foreign_key", "candidate_key" and "value" fields. Id 4 and 6 contains same values in"entity", "attribute_name", "primary_key", "foreign_key", "candidate_key" and "value fields". One entry will be deleted between id 4 and 6 from the "table_semantic". Hence the refinement instance of the table named "table_semantic" has been given in Table VIII.

Some predefine postulates have been taken for SQL formation from one of our research article has been described in [6]. The desired result retrieving query in SQL can be given by SELECT attribute 1, attribute 2, attribute 3… attribute n FROM entity 1 (table 1), entity 2 (table 2), entity 3 (table 3)… entity n (table n) WHERE condition 1 and condition 2 and condition 3… and condition n. There are few clauses are fixed any retrieving query. These are SELECT, FROM and WHERE. The proposed system has to determine the correct possible attribute(s), entities and condition(s). Predefine postulates have been described below as well as conditional flowchart has been given in Fig. 3.

TABLE VIII. REFINEMENT TABLE OF SEMANTIC TABLE

id	entity	attribute_name	primary_key	foreign_key	candidate_key	value
1	table_hospital	hos_add	hos_id			আসানসোল(asansol)(propoer Noun)
2	table_hospital	hos_name	hos_id			
3	table_department	dept_name	dept_id	hos_id		মনোরোগ(mnor[2]og)(mental disease)
4	table_doctor	doc_specialist	doc_id	hos_id	dept_id	মনোরোগ(mnor[2]og)(mental disease)

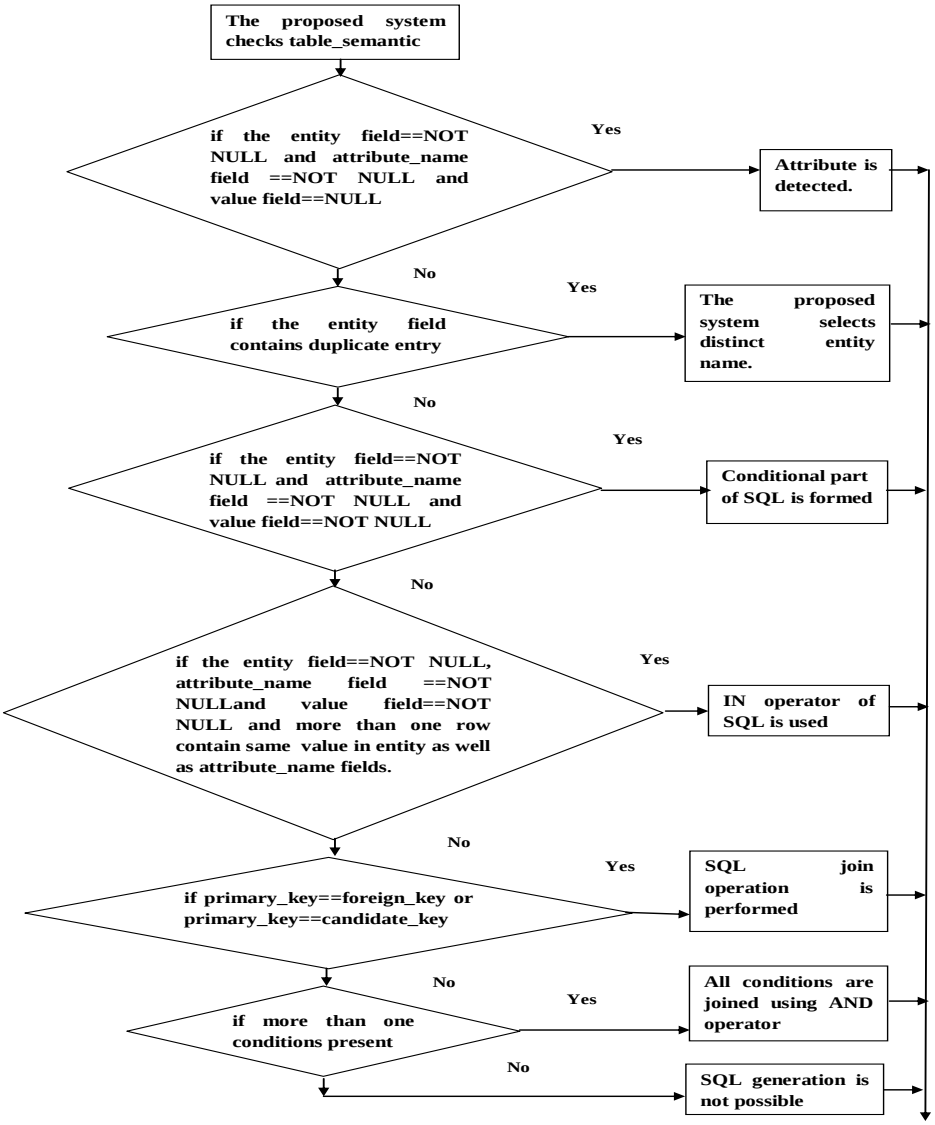

Fig. 3. Conditional Flowchart for SQL Rules Generation.

Postulate i: The proposed system predicts attribute if the "entity" field, "attribute_name" field and "value" field of the table named "table_semantic" contain NOT NULL, NOT NULL, NULL value respectively, then whatever attribute present in the "attribute_name" field is treated as attribute. In case such type of condition does not occur, then the proposed system considers all attributes i.e. denoted by "*". Finally this attribute is appended by "." operator with their corresponding entity field value of that row.

Postulate ii: The system predicts entity name from entity field value(s) of the table named "table_semantic". For duplicate entry the system selects distinct entity field value.

Postulate iii: The proposed system predicts condition if the "entity" field, "attribute_name" field and "value" field of the table named "table_semantic" does not contain NULL, NULL, NULL value respectively, then whatever attribute present in the "attribute_name" field will be treated as condition. This attribute will be appended by "." operator with their corresponding entity field value followed by "=" symbol and value of the field value of that row in the semantic table.

Postulate iv: The proposed system determines IN clause in condition if the "entity" field, "attribute_name" field and value field of the table named "table_semantic" does not contain NULL, NULL, NULL value respectively. More than one row in the table named "table_semantic", the "entity" field as well as corresponding "attribute_name" field contains same value means that particular attribute of that entity has a list of values. Then the attribute will be appended by "." operator with their corresponding "entity" field value followed by IN clause and a list of values will be placed within opening and closing parenthesis separated by ",".

Postulate v: The "primary_key" field value of one entity matches with "foreign_key" or "candidate_key" of other entity that means joining occurs.It is a property of relational database. The proposed system appends matched "primary_key" field value by "." with corresponding "entity" field value followed by "=" symbol and matched "foreign_key" or "candidate_key" value append by "." with corresponding "entity" field value.

Postulate vi: The proposed system combines all conditions using AND operator if more than one conditions are present.

The proposed system converts the user given Bengali query i.e. আসানসোলে মাথা খারাপের চিকিৎসার জন্য হাসপাতাল কোথায় আছে? into SQL has been given below.

SELECT hos_name FROM table_hospital, table_department, table_doctor WHERE table_hospital.hos_add= 'আসানসোল' AND table_hospital.hos_id = table_department.hos_id AND table_department.dept_id = table_doctor.dept_id AND table_department.dept_name = 'মনোরোগ' AND table_doctor.doc_specialist = 'মনোরোগ'.

Component 8: SQL executed by the proposed system.

The proposed system will execute the SQL that will be generated from user given query i.e. given in Bengali language. After execution of the SQL, the expected result will be generated. The response of the user query1 has been given in Fig. 4.

Conversion of Natural Language Query to SQL:

SELECT hos_name FROM table_hospital, table_department, table_doctor WHERE table_hospital.hos_add = "আসানসোল" AND table_hospital.hos_id = table_department.hos_id AND table_department.dept_id = table_doctor.dept_id AND table_department.dept_name = "মনোরোগ" AND table_doctor.doc_specialist = "মনোরোগ"

The Generated Response:

hos_name
ই এস আই

Fig. 4. Response of the Query 1 in Bengali.

VI. Modified Bengali WordNet and Health Information Databases

The data repository of XBLQPS is made up of a modified Bengali WordNet database and a health information database. The modified Bengali WordNet database has been developed with help of Bengali WordNet database as referred in [3]. Table IX, Table XI, Table XII and Table XIII has been kept same as the Bengali WordNet whereas, Table X has been modified by introducing two fields named meaning and "possible_attribute" by the authors. The Health information database contains other tables named "table_inflectional_part","table_entity_attribute_sensing","table_semantic","table_hospital","table_department" and "table_doctor" which have been introduced and developed by the authors. The table named "table_semantic", "table_hospital", "table_department" and "table_doctor" have been taken from our research article referred in [6]. The proposed system thus has modified the original WordNet database of [3] and shall now be termed as "Modified Bengali WordNet DB". Each word in the Modified Bengali WordNet DB contains its parts of speech definition introduced manually as according to the largest probability of its use. We have not referred to any standard Bengali POS Tagger but to our own prepared POS definitions in Modified Bengali WordNet DB. The Modified Bengali WordNet DB and Health information database consist of a set of tables as follows:

Structure of table "table_word"

The table named "table_word" consists of two fields' "word_id" and "word_name". The field "word_id" is primary key field whereas "word_name" field contains Bengali

word.Data types and sizes of these two fields are int(10) and varchar(100) respectively. The instance of the table "table_word" has been given in Table IX.

TABLE IX. STRUCTURE OF TABLE "TABLE_WORD"

word_id	word_name
7	বিভাগ(bibʱag) (department)
16	অপথ্যালমোলজি (ɔptʰɛlmolodʒi) (ophthalmology)
23	মাথাখারাপ(matʰakʰarap) (mental disease)

Structure of table "table_synset"

The table named "table_synset" consists of "synset_id", "parts_of_speech", "gloss_concept_defination", "example_sentence", "meaning" and "possible_attribute". The "synset_id" is the primary key field of this table. The "parts_of_speech" field contains possible parts of speech value of the Bengali word which is stored in table "table_word". The "gloss_concept_defination", "example_sentence" and "meaning" field contain definition, example of sentence and meaning of above mention Bengali word respectively. The "possible_attribute" field contains likelihood relationship of a particular Bengali word with medical domain. If a "possible_attribute" field does not contain any value that means it has NULL value.Data types and sizes of these six fields are int(10), varchar(50), varchar(150), varchar(200), varchar(300) and varchar(150) respectively. The instance of the table "table_ synset" has been given in Table X.

TABLE X. STRUCTURE OF TABLE "TABLE_SYNSET"

synset_id	parts _of_ speech	gloss _concept_ definition	example_sentence	meaning	possible_attribute
2	বিশেষ্য(biʃeʃo)(noun)	একটি প্রতিষ্ঠানের বিশেষ অংশ	রবি রয় অস্থিচিকিৎসা বিভাগে রডাক্তার		dept_name
5	বিশেষ্য(biʃeʃo)(noun)	চক্ষুরোগ সমন্ধীয় চিকিৎসা	চক্ষুরোগে আক্রান্ত ব্যক্তি রাত্রে চশমা ব্যবহার করেন	চক্ষুবিজ্ঞান	dept_name,doc_specialist
10	ক্রিয়া(kir[2]ia) (verb)	কোন অনিশ্চতার সমন্ধে উদ্বেগ	তুচ্ছ ব্যপার নিয়ে তুমি মাথা খারাপ করোনা	বিরক্ত করা	

Structure of table "table_sense"

The table named "table_sense" has two fields. These two fields are "word_id" and "synset_id". The primary key field consists of both fields. This table is used to map a "word_id" to its corresponding "synset_id". Data types and sizes of these two fields are int(10) and int(10) respectively. The instance of the table "table_ sense" has been given in Table XI.

TABLE XI. STRUCTURE OF TABLE "TABLE_SENSE"

word_id	synset_id
7	2
16	5
23	10

Structure of table "table_hypernym"

The table named "table_hypernym" consists of two fields. These two fields are "synset_id" and "hypernym_id". The "synset_id" field stores the synsetid value of a Bengali word which is stores in the table "table_word" whereas "hypernym_id" field value is nothing but a "synset_id" of a another word. These two words have hypernym and hyponym relationship.Data types and sizes of these two fields are int(10) and int(10) respectively. The instance of the table "table_hypernym" has been given in Table XII.

TABLE XII. STRUCTURE OF TABLE "TABLE_HYPERNYM"

synset_id	hypernym_id
7	3
8	3

Structure of table "table_tree"

The table named "table_tree" consists of two fields. These two fields are "hypernym_id" and "parent_id". The "hypernym_id" is "synset_id" of a Bengali word. The "parent_id" is parent word id of a hypernym word. This table maintains the hierarchical relationship between hypernym and hyponym word. The parent_id field value contains zero that represents root word of the tree. Data types and sizes of these two fields are int(10) and int(10) respectively. The instance of the table "table_tree" has been given in Table XIII.

TABLE XIII. STRUCTURE OF TABLE "TABLE_TREE"

hypernym_id	parent_id
3	0

Structure of table "table_entity_attribute_sensing"

The table named "table_entity_attribute_sensing" consists of "id", "entity", "attribute_name", "primary_key", "foreign_key" and "candidate_key". The "id" field is the primary key field of this table. The "entity" field contains name of participating entities in medical domain. The "attribute_name" field stores attribute name, "primary_key" field contains primary key, "foreign_key" contains foreign key and "candidate_key" stores candidate key value of corresponding entity. The sometimes foreign key and candidate key may not exist, that time these key fields contain NULL value.The "table_entity_attribute_sensing" is an independent table. This tablehas been used to map entity from attribute, or primary key, foreign key, candidate key from entity.Data types and sizes of these six fields are int(10), varchar(50), varchar(50), varchar(50), varchar(50) and varchar(50) respectively. The instance of the table "table_entity_attribute_sensing" has been given in Table XIV.

TABLE XIV. STRUCTURE OF "TABLE_ENTITY_ATTRIBUTE_SENSING"

id	entity	attribute_name	primary_key	foreign_key	candidate_key
1	table_hospital	hos_name	hos_id		
2	table_hospital	hos_add	hos_id		
3	table_hospital	hos_district	hos_id		

Structure of table "table_inflectional_part"

The table named "table_inflectional_part" consists of "id" and "inflectional_part". The id field is primary key field of this table. The "inflectional_part" field stores inflectional part of the Bengali word. Data types and sizes of these two fields are int(10) and varchar(50) respectively. The instance of the table "table_inflectional_part" has been given in Table XV.

TABLE XV. STRUCTURE OF "TABLE_INFLECTIONAL_PART"

id	inflectional_part
1	ি(i)
2	ে(e)
3	ের(er[2])

Structure of table "table_semantic"

The table named "table_semantic" includes id, entity, "attribute_name", "primary_key", "foreign_key", "candidate_key" and "value" fields. The "id" field is the primary key of this table. The entity name, corresponding attribute name, primary key, foreign key, candidate key and value are held by fields named"entity", "attribute_name", "primary_key", "foreign_key", "candidate_key" and "value" respectively. The "table_semantic" is an independent table. This table has been used to construct the SQL.Data types and sizes of these seven fields are int(10), varchar(50), varchar(50), varchar(50), varchar(50), varchar(50) and varchar(100) respectively. The instance of the table "table_semantic" has been given in Table XVI.

TABLE XVI. STRUCTURE OF "TABLE_SEMANTIC"

id	entity	attribute_name	primary_key	foreign_key	candidate_key	value
1	table_hospital	hos_add	hos_id			আসানসোল(asansol) (propoer Noun)
2	table_hospital	hos_name	hos_id			
3	table_department	dept_name	dept_id	hos_id		মনোরোগ(mnor[2]og)(mental disease)

Structure of table "table_hospital"

The table named "table_hospital" includes "hos_id", "hos_name", "hos_add", "hos_district", and "hos_state" fields. The "hos_id" field is the primary key of this table. This "hos_id" field used to identify uniquely each entity instance of the table. The hospital name, hospital address, district name and state name where hospital is situated, are held by fields named "hos_name", "hos_add", "hos_distrct", and "hos_state" respectively.Data types and sizes of these five fields are int(10), varchar(100), varchar(100), varchar(100) and varchar(100) respectively. The instance of the table named "table_hospital" has been given in Table XVII.

TABLE XVII. STRUCTURE OF "TABLE_HOSPITAL"

hos_id	hos_name	hos_add	hos_district	hos_state
1	মুর্শিদাবাদজেলাহাসপাতাল (mur[2]ʃidabaddʒelaɦaspatal) (Murshidabad District Hospital)	লালগোলা(lalgola) (Lalgola) (proper noun)	মুর্শিদাবাদ (mur[2]ʃidabad) (Murshidabad) (proper noun)	পশ্চিমবঙ্গ(pʃtʃimbɔŋg) (West Bengal) (proper noun)
8	হাওড়াজেলাহাসপাতাল (ɦaoɽadʒelaɦaspatal) (Howrah District Hospital)	আমতা(amta) (Amta) (proper noun)	হাওড়া (ɦaoɽa) (Howrah)(proper noun)	পশ্চিমবঙ্গ(pʃtʃimbɔŋg) (West Bengal) (proper noun)

Structure of table "table_department"

The table named "table_department" includes "dept_id", "dept_name", and "hos_id"fields.Data types and sizes of these three fields are int(10), varchar(100) and int(10) respectively. The instance of the table "table_department" has been given in Table XVIII.

TABLE XVIII.STRUCTURE OF "TABLE_DEPARTMENT"

dept_id	dept_name	hos_id
70	নবজাতক(nbdʒatk) (neonate)	7
170	মনোরোগ(mnor[2]og) (psychiatry)	12

Structure of table "table_doctor"

The table named "table_doctor" includes "doc_id", "doc_name", "doc_qualification", "doc_ specialist", "hos_id" and "dept_id" fields. Data types and sizes of these six fields are int(10), varchar(100), varchar(100), varchar(100), int(10) and int(10) respectively. The instance of the table "table_doctor" has been given in Table XIX.

TABLE XIX. STRUCTURE OF "TABLE_DOCTOR"

doc_id	doc_name	doc_qualification	doc_specialist	hos_id	dept_id
1000	কেয়া গরাই(keagr[2]ai) (Keya Gorai)	এম.বি.বি.এস. (M.B.B.S.)	অপথ্যালমোলজি(ɔpthɛlmolodʒi) (Ophthalmologist)	1	10
1010	শমীতা দাশগুপ্ত(ʃmitadaʃgupta) (Shamita Dasgupta)	এম.ডি. (M.D.)	অস্থি(ɔsthi) (Orthopaedist)	2	20

VII. Tools and International Phonetic Alphabet (IPA) Notation used

Software tools like HTML, PHP, MySQL and Avro Bengali keyboard has been used to developed the XBLQPS. The HTML has been used to develop the web pages structure. A sever side scripting language i.e. PHP has been used to provide customize interface as well as process the user request. MySQL is used as back end database to store the data. The IPA notation for Bengali language has been taken from the website https://en.wikipedia.org/wiki/Help:IPA/Bengali.

VIII. Time Complexity

The time complexity has been calculated for algorithmic steps of the proposed system.

Step 1: Login into the system and submit the query in Bengali language.

Let, as an example query আসানসোলে মাথা খারাপের চিকিৎসার জন্য হাসপাতাল কোথায় আছে?. The time complexity has been calculated on the proposed algorithm using the above example query. The time has been calculated in each algorithmic step that is given below.

Step 2: Tokenization

Let, there be n number of token(s) in user given query has been given in Fig. 5.

Time taken for tokenization of n token(s) = n unit time.

The number of tokens will be generated from above example query has been given in Table XX.

Here for this query time taken for tokenization=8 unit time because number of token(s) i.e. n=8.

Step 3: Root word extraction

Let, there be p numbers(s) of inflection part in the table "table_inflectional_part", and also assume that q number(s) of token(s) which are matched with value of "inflectional_part" field.

∴ Time taken= p×q unit time.

Here p=3, q=3.

Therefore, time taken=3×3=9-unit time.

Step 4: POS tagging

Let, there be r number(s) of "word_name" in the "table_word".

Let, there be r number of "word_id" and "synset_id" in the table "table_sense".

Let, there be r number(s) of parts of speech in the table "table_synset".

∴Time taken = (p×r+r×r+r×r) unit time.

Here r=3 for "table_word".

r=3 for "table_sense".

r=3 for "table_sysnset".

∴Time taken= (3×3+3×3+3×3) unit time=27-unit time.

Step 5: Pattern generation of noun phrases

Let, n number(s) of token(s) are participating for pattern generation where the token occurrence orders are to be maintained.

∴Time taken=n+(n-1) +(n-2)+⋯+1-unit time.

=n/2 (n+1)

There are 11 numbers of patterns which are already generated in Table V.

Therefore, time taken = 11 unit time.

Step 6: Sense extraction of pattern(s) using LESK algorithm and semantic analysis.

Time taken for unambiguous pattern={n+(n-1)+(n-2)+⋯+1}(r+r+r) unit time.

=n/2 (n+1)(r+r+r)

Let, there be u number(s) of similar "word_id" in the "table_sense".

Let, v number(s) of "synset_id" contains "possible_attribute" field value among u number(s) of similar "synset_id".

Let, x number(s) of values exists "attribute_name" in the table "table_entity_attribute_sesing".

Let, y number(s) of instance(s) have been inserted in the table "table_semantic".

Time taken for ambiguous pattern={ n+(n-1)+(n-2)+⋯+1}(r+r×u+u×r+v×x+v×y) unit time.

Time taken for unambiguous patterns =8(3+3+3)=72 unit time.

u=number of similar "word_id" in the "table_sense" =3.

v=3, x=3, y=6.

∴Time taken for ambiguous patterns= 8(3+3×3+3×3+3×3+3×6) unit time=384 unit time.

Fig. 5. Token(s) of user given Query.

TABLE XX. Token of user given Query.

Number of tokens	1	2	3	4	5	6	7	8
Token of example query	আসানসোলে (asansole)	মাথা (mat^{h}a)	খারাপের (k^{h}araper)	চিকিৎসার (ʧikitsar)	জন্য (ʤno)	হাসপাতাল (ɦaspatal)	কোথায় (kot^{h}aę)	আছে (atʃ^{h}e)

Step 7: SQL formation

Let, time taken for SQL formation=z unit time.

6 postulates have been applied on 4 rows of "table_semantic" to form SQL.

∴Time taken =(4×6) unit time =24 unit time.

Step 8: SQL executed by the proposed system

Let, time taken for SQL execution=t unit time.

∴Time complexity =f(n,p,q,r,t,u,v,x,y,z)

=n+p×q+(p×r+r×r+r×r)+n+(n-1)+(n-2)+⋯+1+{n+(n-1)+(n-2)+⋯+1}(r+r+r)+{n+(n-1)+(n-2)+⋯+1}(r×u)+r×u+u×r+v×x+v×y+z+t

∴f(n)≅n+n×n+(n×n+n×n+n×n)+n+(n-1)+(n-2)+⋯+1+{n+(n-1)+(n-2)+⋯+1}(n+n+n)+{n+(n-1)+(n-2)+⋯1}n+n×n+n×n+n×n+n×n+n+n

$$= n + n^2 + 3n^2 + \frac{n}{2}(n+1) + \left\{\frac{n}{2}(n+1)\right\}(3n) + \left\{\frac{n}{2}(n+1)\right\}n + n^2 + n^2 + n^2 + n + n = n + 4n^2 + \frac{n^2}{2} + \frac{n}{2} + \frac{3n^2}{2}(n+1) + \frac{n^2}{2}(n+1) + 4n^2 + n + n$$

$$= n + 8n^2 + \frac{n^2}{2} + \frac{n}{2} + \frac{3n^3}{2} + \frac{3n^2}{2} + \frac{n^3}{2} + \frac{n^2}{2} + n + n$$

$$= O(n^3)$$

Time taken for SQL execution t=11 unit time.

∴Time complexity algorithm.

= (8+9+27+11+72+384+24+11) unit time.

=546 unit time.

IX. Precision, Recall and Accuracy Analysis of the XBLQPS

Example of Precision, recall and accuracy analysis of the proposed XBLQPS.

111 numbers of queries have been submitted into the proposed system by the user. Four different types of queries based on their output have been given as examples. The output has been classified into "expected output" and "select output". "expected output" and "select output" have been further categorized into positive and negative. The Confusion Matrix [23], [24] for the proposed system has been given in Table XXI.

TABLE XXI. Confusion Matrix

Expected output Vs. Select output	Select output-Positive	Select output-Negative
Expected output-Positive	True positives (TP)-46	False positives (FP)-16
Expected output-Negative	False negatives (FN)-17	True negatives (TN)-32

Some categorized sample queries has been given below as an example.

TP- আসানসোলে মাথা খারাপের চিকিৎসার জন্য হাসপাতাল কোথায় আছে?

TN- বাকুড়া জেলাহাসপাতালে অ্যাম্বুলেন্সের ব্যবস্থা আছে?

FP- সুমিত রায়ের শিক্ষাগত যোগ্যতা কি?

FN-বাকুড়া হাসপাতাল

Precision = TP/(TP+FP)=46/(46+16)=46/62=0.74

Our system has a precision of 0.74 - in other words, when it predicts a query is fetched, it is correct 74% of the time.

Recall = TP/(TP+FN)=46/(46+17)=46/63=0.73

Our system has a recall of 0.73 - in other words, it correctly identifies 73% of all fetched queries.

Accuracy=(TP+TN)/(TP+FP+FN+TN)=(46+32)/(46+16+17+32)=78/111=0.70

Our system has an accuracy of 0.70 - in other words, it gives 70% correct predictions of fetched queries.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)=(2*0.73*0.74)/(0.73+0.74)=1.08/1.47=0.73

Our system has F1 score is 0.73 – in other words, overall measure of our system's accuracy is 73%.

X. Future Work

The XBLQPS uses a simplified version of the LESK algorithm. It depends on the maximum number of the common word(s) between word gloss and the current context of the word. As the gloss may be limited for a domain in the proposed Bengali WordNet, it may happen that the proposed system sometimes fails to disambiguate an ambiguous word. Our proposed system uses a manually annotated parts of speech method for the detection of parts of speech of a word which is based on the largest probability of its use. The proposed system needs to be improved by using a modified version of the LESK algorithm and automated POS tagging method for better handling natural language queries with ambiguous words and this can be marked as a future work of this proposed system.

XI. Conclusion

The XBLQPS uses a simplified version of the LESK algorithm. It depends on the maximum number of the common word(s) between word gloss and the current context of the word. As the gloss may be limited for a domain in the proposed Bengali WordNet, it may happen that the proposed system sometimes fails to disambiguate an ambiguous word. Our proposed system uses a manually annotated parts of speech method for the detection of parts of speech of a word based on the largest probability of its use. In the future, the proposed system needs to be improved by using a modified version of the LESK algorithm and automatic POS tagging method for better efficiency wherein the gloss shall be enhanced to

accommodate several synonymous applications of a word in that particular domain. Moreover, we may enhance the proposed system to work with unstructured databases. Retrieving data from a relational database management system from a Bengali language query containing the ambiguous word is a challenging task. The proposed XBLQPS will be able to handle a domain-specific Bengali language query containing an ambiguous word which will be helpful even for a naive user. The naïve user can access the database without knowledge of for-mal language i.e. SQL (Structured Query Language). The XBLQPS is an enhancement and advanced form of BLQPS of the work discussed in [6]. The XBLQPS incorporates the handling of ambiguous Bengali words using the LESK algorithm and a modified WordNet in Bengali. The time complexity, precision, recall, accuracy and F1 score have been analyzed for our proposed system. The time complexity of XBLQPS is $O(n^3)$ as compared to the time complexity of BLQPS of [6] which is $O(n^4)$ which shows that the time efficiency of XBLQPS is better than BLQPS.

ACKNOWLEDGMENT

This study was conducted at the National Institute of Technology (NIT), Durgapur, Research Project Lab, which is part of the Department of Computer Science and Engineering. The authors express their gratitude to the Department of Computer Science and Engineering, NIT, Durgapur, India, for providing educational resources for this research.

REFERENCES

[1] H.K. Azad, and A. Deepak, "A new approach for query expansion using Wikipedia and WordNet," Elsevier, Vol. 492, pp. 147-163, 2019.

[2] N.S. Dash, P. Bhattacharyya, and J.D. Pawar, "The WordNet in Indian Languages," Springer, pp. 243-260, 2017.

[3] F. Faruqe, and M. Khan, "Bwn-a software platform for developing bengali wordnet," Innovations and Advances in Computer Sciences and Engineering., Springer, pp. 337-342, 2010.

[4] S. Madonsela, "African Wordnet as a tool to identify semantic relatedness and semantic similarity," South African Journal of African Languages, Taylor & Francis, vol. 39, no. 2, pp. 185-190, 2019.

[5] O. Dongsuk, S. Kwon, K. Kim, Y. Ko, "Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph," Proceedings of the 27th international conference on computational linguistics, pp. 1-12, 2018.

[6] K.P Mandal, P. Mukherjee, A. Chattopadhyay, B. Chakraborty, "A novel Bengali Language Query Processing System (BLQPS) in medical domain," Intelligent Decision Technologies, IOS Press, vol. 13, no. 2, pp. 177-192, 2019.

[7] G. Huilin, G. Tong, W. Fan, M. Chao, "Bidirectional Attention for SQL Generation," 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 676-682, 2019.

[8] P. Sharma, N.J.E. Joshi, Technology, and A.S. Research, "Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet," Engineering, Technology and Applied Science Research, vol. 9, no. 2, pp. 3985-3989, 2019.

[9] A. Sau, T.A. Amin, N. Barman, A. R. Pal, "Word Sense Disambiguation in Bengali Using Sense Induction," 2019 International Conference on Applied Machine Learning (ICAML), IEEE, pp. 170-174, 2019.

[10] M.M. Rahman, S.A. Khan, and K.A. Hasan. "Word Sense Disambiguation by Context Detection," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), IEEE, pp. 1-6, 2019.

[11] H. Walia, A. Rana, and V. Kansal. "Case based interpretation model for word sense disambiguation in Gurmukhi," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp. 359-264, 2019.

[12] M.S. Kaysar, M. A. B. Khaled, M. Hasan, M. I. Khan, "Word sense disambiguation of Bengali words using FP-growth algorithm," 2019 international conference on electrical, computer and communication engineering (ECCE), IEEE, pp. 1-5, 2019.

[13] G. Jain, D. Lobiyal, "Word Sense Disambiguation of Hindi Text using Fuzzified Semantic Relations and Fuzzy Hindi WordNet," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp. 494-497, 2019.

[14] N.S. Dash, "Back to Basics: A Road to Return to Nominal Base through Lemmatization," Proceedings of Abstracts of the 36th International Conference of the Linguistic Society of India (ICOLSI-36), pp. 1-34, 2014.

[15] S. Basuki, A. S. Kholimi, A. E. Minarno, F. D. S. Sumadi, M. R. A. Effendy, "Word Sense Disambiguation (WSD) for Indonesian Homograph Word Meaning Determination by LESK Algorithm Application," 2019 12th International Conference on Information & Communication Technology and System (ICTS), IEEE, pp. 8-15, 2019.

[16] Y. Heo, S. Kang, J.J.I.A. Seo, "Hybrid sense classification method for large-scale word sense disambiguation,"IEEE, vol. 8, pp. 27247-27256, 2020.

[17] C. Lachichi, C. Bendiaf, L. Berkani, A. Guessoum, "An Arabic WordNet enrichment approach using machine translation and external linguistic resources," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), IEEE, pp. 1-6, 2018.

[18] Q.-P. Nguyen, A.-D. Vo, J.-C. Shin, P. Tran, C.-Y. Ock, "Korean-vietnamese neural machine translation system with korean morphological analysis and word sense disambiguation,", IEEE, vol. 7, pp. 32602-32616, 2019.

[19] A. Jabbar, S. Iqbal, A. Akhunzada, Q. Abbas, "An improved Urdu stemming algorithm for text mining based on multi-step hybrid approach," Journal of Experimental & Theoretical Artificial Intelligence, Taylor & Francis, vol. 30, no. 5, pp. 703-723, 2018.

[20] Q. Zhou, Y. Meng, "combination of Semantic Relatedness with Supervised Method for Word Sense Disambiguation," 2019 International Conference on Asian Language Processing (IALP), IEEE, pp. 142-147, 2019.

[21] M. Biswas, M.M. Hoque, "Development of a Bangla Sense Annotated Corpus for Word Sense Disambiguation," 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, pp. 1-6, 2019.

[22] A. Haque, M.M.J.I. Hoque, "Bangla word sense disambiguation system using dictionary based approach," pp. 1-6, 2016.

[23] A. Jakka, J.J.I.J.I.T.E.E. Vakula Rani, "Performance Evaluation of Machine Learning Models for Diabetes Prediction," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 11, pp. 1976-1980, 2019.

[24] A. Santra, J. Christy, "Genetic algorithm and confusion matrix for document clustering," International Journal of Computer Science Issues, vol. 9, no. 1, pp. 322, 2012.

[25] K. JanakiRaman, K. Meenakshi, "Automatic Text Summarization of Article (NEWS) Using Lexical Chains and WordNet," International Journal of Advanced Science and Technology, Vol. 29, no. 4, pp. 3242-3258, 2020.

[26] A. Alkhatlan, J. Kalita, A. Alhaddad, "Word Sense Disambiguation for Arabic Exploiting Arabic WordNet and Word Embedding," The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), Elsevier, Vol. 142 pp. 50-60, 2018.

[27] A. H. Aliwy, A. R. Abbas, "Improvement WSD Dictionary Using Annotated Corpus and Testing it with Simplified Lesk Algorithm," Fifth International conference on Computer Science and Information Technology, pp. 89-97, 2015.

AUTHORS' PROFILE

Kailash Pati Mandal his obtained Bachelor of Technology in Computer Science and Engineering from Bengal College of Engineering and Technology, Durgapur, India in 2006. In 2011, he received Master's degree in Computer Science and Engineering from the Jadavpur University, Kolkata, India. He is currently a part-time Researcher in Computer Science and Engineering in the field of Natural Language Processing at the National Institute of Technology (NIT), Durgapur, India.

Prasenjit Mukherjee obtained his PhD in Computer Science and Engineering from National Institute of Technology (NIT), Durgapur, India under Visvesvaraya PhD scheme Program, Ministry of Electronics and Information Technology (MeitY), Govt. of India. His research interests include machine learning, deep learning, natural language processing, knowledge engineering, and database management systems. He has more than 15 international publications.

Atanu Chattopadhyay received the M.Sc degree in Applied Mathematics and Computing from Indian Institute of Technology (Indian School of Mines), Dhanbad, India. He is presently a fulltime Lecturer in the Department of BBA(H) & BCA (H), Deshabandhu Mahavidyalaya, Chittaranjan, West Bengal, India. His research interest includes Mathematical Analysis and Natural Language Processing.

Baisakhi Chakraborty obtained her PhD in Computer Science and Engineering from National Institute of Technology, Durgapur, India in 2011. Her research interests include knowledge systems, knowledge engineering and management, database systems, data mining, natural language processing, and software engineering. She has several researchers under her direction. She has more than 30 international publications. She has a decade of industry experience and 17 years of academic experience.