IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. I (Jan – Feb. 2016), PP 01-08
www.iosrjournals.org

NLP for Automated Conceptual Data Modeling

[1]Suraj A. Jogdand,[2]Mr. Pramod B. Mali,
[1]*PG Student,Department of Computer Engineering.Smt. KashibaiNavale College of Engineering, Vadgaon Bk, Pune, Maharashtra, India*
[2]*Assistant Professor,Department of Computer Engineering.Smt. KashibaiNavale College of Engineering, Vadgaon Bk, Pune, Maharashtra,India*

Abstract:*One of the procedures for obtaining the better understanding of the system by splitting the large system into smaller parts, this procedure is termed as analysis. Manually creating softwareartifacts and building UML models from natural language is a complex job. We represent diagrammatically structure of data by ER diagram. This paper represents the use of heuristics picked up from the syntactic and semantic learning for the automatic formation of database plan, which leads us to the identification and methodology to generate Entity-Relationship(ER) through NLP.*

The use of both syntactic and semantic heuristics is used as the system to obtain the critical ER parts, components, qualities and associations from the specifications. In this method, we take English as input from the client and our framework at first create summary of that few paragraphs utilizing code quantityprinciple. The resultant summary is utilized in our automatic ER chart generation framework. Atlast a XMI record is made. This file can be pictured in any of the UML modeling tools. Experimental Results on the usage of the semantic heuristics demonstrate that these may help to further upgrade the outcomes in the modified differentiation of the ER segments.

Keywords: *Entity-Relationship Model,E-R Diagram, Data Modeling, XML, POS, Tokenization, NLP.*

I. Introduction

AnEntity-Relationship model is a purposeful system for representing and describing a business process. The strategy is shown as parts that are associated with each other by relationshipswhich articulate the conditions and requirements amongthose, for example, one fusing may be differentiated with zero or more apartments, yet one apartment must be put in one building. Entities may have distinct properties which portray them. Diagramsgenerated to identify with these elements, attributes, and connections graphically are called entityrelationship graphs. An E-R model is ordinarily actualized as a database. By virtue of a social database, which stores data in tables, every column of every one table addresses one event of an entity. Some data fields in these tables point to documents in distinct tables, such pointers address the relationship.

The natural language English has various ambiguities. The most remarkable among them is that diverse individuals have particular interpretations of the same statement. These ambiguities make an issue in manual generation of E-R chart. E-R frameworks are fundamental at the present time programming progression as mistakes in the ER diagram can get reflected in the last item. Subsequently it is fundamental to make a correct ER graph created at this stage can be difficult to modify later on. Analyzing requirements and creating the item curious to assemble an ER graph is a confounded task. It is obliged to regularly create the ERD of a structure with minimum human associations. A robotized tool can be created which will remove the appreciativedata to create programming remainder which will be further utilized to build up the ER diagram of the given English paragraphs.

Examination is a basic stage in development of software. The UML models gatheredduring the examination stage help in the complete process of software development. This paper proposes a framework for generating ER Diagram from the created summary i.e. communicated in English language. Initially a client necessity is to enter the input paragraphs. Theseinput paragraphsare changed over into transitional semi-formal representation termed as semantics vocabulary and rules. At this point synopsis gets changed over into transitional semi-formal representation called as Semantics Vocabulary and rules. The Semantics Vocabulary Rules is an approved standard of the Group of Object Management. Database setup is a methodology of making a consistent data model for a specific database. Entity relationship showing, which is an irregular state sensible model planned to empower database outline, can beoverwhelming undertaking to both understudies and fashioners alike in light of the fact that of its extraordinary nature and subtle element. Much research has attempted to apply natural language generating in removing data from requirements points of interest with the plan databases.

Regardless, focus on the improvement and usage of heuristics to help the improvement of sensible databases from natural language has been uncommon. The SBVR (Semantic Business vocabulary and Rule)

characterizes the vocabulary and rules for reporting the semantics of business vocabularies, business actualities and business rules. The SBVR representation of the issue explanation is given to the POS (Part of discourse) labeling stage where every statement is labeled with the suitable parts of discourse. This labeled record is given to the occasion extraction stage where occasions are separated Events are Entity sort, Substance, Relationship sort, Attribute sort, Attribute for Entity, Property for relationship. The removed occasions are mapped to the components of the ER graph. At last a XMI i.e. nothing however XML Metadata Interchange document is created which can be foreign made in any of the UML demonstrating tool to see the created ER chart.

II. Literature Survey

For generating the entity relationships there is need of entity and relationship in paragraph. For that in this paper framework uses code quality guideline [1] for generating the summary. Programmed contented summarization by brushing the Information Retrieval (IR) methodologies with the semantic models code sum, memory and respect for get the significantsentences [2]. In our procedure we are refining the most basic information from the source to get the genuineconcept in the dense structure. We have initially emptied the repetition of the information and we have paced the sentences concentrated around the linguistic principles code quantity, memory and consideration.

The methodology begins by analyzing a plain entered substance data containing a necessities determination of a database problem in English. Therefore, a parser is obliged to parse the English sentences to get their grammatical feature [2] marks before further taking care. Grammatical feature marking selects each one saying in a data sentence it's fitting grammatical form, for example, verb and determiner to mirror the announcement's syntactic class. Grammatical feature short structure and its importance. Grammatical feature labeling is only discovering each word placein sentence. In Natural language processing POS is imperative for substantially more processing.

The work like DMG [3] gives a premise to the improvement of novel heuristics connected in ER-Converter. DMG is a rule based diagram instrument i.e. outline tool which keeps up standards and heuristics in a couple of information bases. A parsing algorithm which becomes acquainted with a language structure and a vocabulary is planned to meet the essentials of the tools. During the parsing stage, the sentence is parsed by recuperating key information from the semantic utilization, spoken to by syntactic standards and the lexicon. The parsing results are changed further on by rules and heuristics which set up a relationship in the middle semantic and arrangement learning. The DMG needs to connect with the customer if an announcement does not exist in the vocabulary or the information of the mapping rules is indistinct. The linguistic structures are then changed by heuristics into EER ideas. Regardless of the way that DMG proposed incalculable to be used as a part of the change from regular dialect to EER models, the device has not yet been made into a commonsense system.

In [4], much workhas tried to apply trademark dialect in differentiating data from necessities particulars or dialog sessions with makers with the arrangement to plan databases. Dialog tools [3] is a data based device associated with the German language for making anoutline framework of an Enhanced Entity-Relationship (EER) model. This instrument is a bit of a greater database layout structure known as RADD which involves distinct portions that structure a complex device.

E-R generator [5] is a substitute rule based structure that makes E-R models from natural language determinations. The E-R generator contains two sorts of rules: specific standards associated with semantics of a couple of words in sentenceand bland chooses that perceive substances and associations on the reason of the smart sign of the sentence and on the reason of the components and associations being worked on. The learning representation structures are building by a naturallanguageunderstandersystem which uses a semantic interpretation approach. There are circumstances in which the system needs help from the customer with a particular final objective to determine ambiguities, for example, the association of characteristics and deciding anaphoric references.

In [6], a methodology of making ER segments characteristically from natural language particulars using heuristics. Semantic heuristics are proposed to be utilized in conjunction with the syntactic heuristics to upgrade the accuracy of the outcomes in making the ER parts from regular naturallanguage particulars. The dedication made can be joined in zones, for example, some bit of the space model of a smart coaching structure, proposed to help in the learning and training of databases and different applications of NLP for database plan.

In [7], a novel procedure of making an interpretation of natural language to SBVR business standards. Regularly, business standard master needs to physically make a several business administers in a Natural language (NL) and after that physically translate NL determination of each and every one of fundamentals in a particular principle, for instance, SBVR, as required. This paper proposes an automated methodology that characteristically translates the Natural lingo (NL) (i.e. English) particular of business rules to SBVR Semantic Business Vocabulary and Rules standards. These methodologies use a rule based count for healthy semantic

investigation of English, and make SBVR standards. The fundamental methods that we are taking after are tokenization, sentence splitting, parts of speech tagging and morphological investigation.

In [8], Automatic evaluation of sequence diagram they analyze two regions of examination. To begin with, they review the way of the errors made by students when drawing sequence diagram and how that information has affected the setup of a learning device. Second, they discuss the adequacy of his modified checking procedure when joined with sequence diagrams. They close with a discussion of how the two strands of this research can be joined to give anextensive alteration device to help students construct right sequence diagrams.

In [9], they uncover a procedure to the programmedinterpretation ofdiagram assembled arranged in light of a 5-stage framework. The paper depicts procedure to subsequently assessing diagram assembled outlines and reports in light of a couple of investigations into the programmed inspecting of understudy graphs. This paper described both proposed framework general procedure for examining at detached diagrams and how they have used our advancement as a part of particular in set of softwaretools planned for learning and assessing graph based charts. Customary samples of graph based diagrams are entity relationship diagrams, Unified Displaying Language charts, natural flow diagrams and chemical structure diagrams. In the work discussed here, they have used entity relationship diagram as a model of graph based charts.

III. Proposed System

The representation of the proposed system is as follows:

3.1System Architecture

The system is distributed in to five modules. The brief description of these modules is as follows:

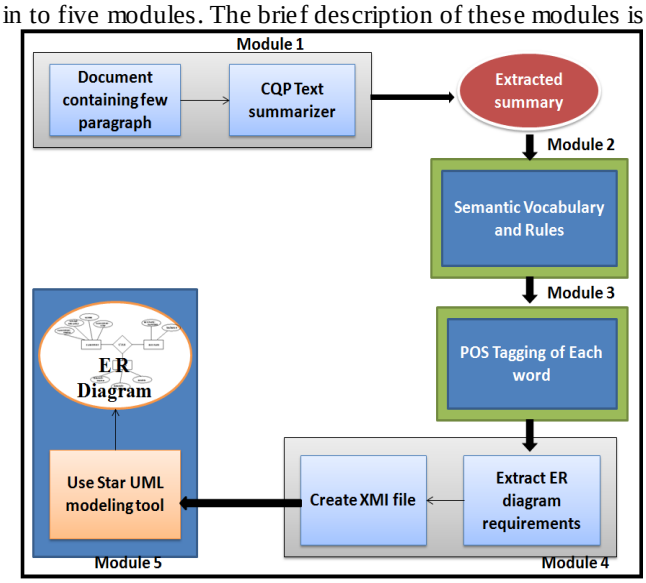


Figure.1 System Architecture

1. **Module 1**: Initially we discussed about the input providing to the system. The natural language textual information is taken as the input to the system. The information contains the few paragraphs.
   - **Pre-processing**: before generation of summary we need to do some process on input text for further process. This process includes tokenization, sentence separation, stop words removing, Stemming, part of speech tagging.
   - **Achieve Dissimilarity**: If two sentence having same meaning (may be one sentence is sub sentence of other sentence) then keep one of them.
   - **Keywords extraction**: Find out most frequent words for sentence ranking
   - **Sentence scoring**: Based on important keyword and noun (CQP) words contain in sentence assign weight to each sentence.
   - **Summary generation**: Based on weight of each sentence select top weighted sentences as summary of all paragraphs or documents.
2. **Module 2:** In this module natural languages specification is translated to SBVR [7] business rules. SBVR is a short type of "Semantic Business Vocabulary and Rules" [7]. To decrease the gap between business expert and IT persons, Semantic Vocabulary and Rules has been presented by Object Administration Group (OMG). This is better method for catching the business necessities in common language. Like structure it is straightforward for people. Because of its higher request of rationale

establishment it is exceptionally easy to machine process. All particular representations and meaning of realities and ideas utilized by an association as a part of course of business are considered as vocabulary. In SBVR a formal presentation under the business impact are considered as rules which are utilized to express the operation of specific business element under particular conditions. Output of this module is summary in semantic vocabulary rule format.

3. **Module 3: Part of Speech Tagging to each word**. In this module the words in a sentence are divided and labeled with specific documentations like NNP for Proper person noun, place or thing, NP for noun phrase, VP for verb phrase, and so on. A parse tree for each one sentence is likewise produced. We utilize this methodology for producing parse tree for a sentence which will be given as input.
4. **Module 4: Extract ER diagram requirement**. For ER diagram construction need to find out relationship, entity and attributes. Common noun as Entity type, Proper noun as Entity, Transitive verb as Relationship Type, Intransitive verb as Attribute type, Adjective as Attribute for entity, adverb as attribute for relationship. Save this output in XMI file.
5. **Module 5: Use star UML modeling tool**. XMI file generated by module 4 is imported in this tool. Then the imported file is opened, which is our expected output i.e. E-R diagram

3.2ALGORITHMS

3.2.1 Summary Generation Algorithm

Input: Multiple Sentence, stop word list file, S num of sentence in final summary.

Output: S num of sentences.

1. Get input paragraph from user
2. Apply stemming
3. Read stop word file
4. Remove stop words from file and select unique keywords
5. Check frequency of each keyword and select those keywords whose frequency is greater than threshold value as final keywords.
6. Assign score of each sentence by CQP method

Select top scored sentences as final summary.

3.2.2 XMI File Generation Algorithm

Input: Summary Sentence, XMI file.

Output: XMI file.

1. Check SBVR rule of each sentence
2. Replace each sentence according to rule.
3. Create POS structure of each file.
4. Select for XMI file element from POS structure for each sentence.
5. Select Noun as entity, verb as relationship, adjective as attribute.
6. Generate XMI file.

3.3 Set Theory

The system S is represented as: S=f HD, VD, TG, ED, DDg

(a) Input Paragraph to the System

Let P is the set of input P=p1, p2,....,pn

Where,

p1, p2,…,pn are the set of inputs.

(b) Text Summarization Process

Let TS is the set of processes TS=sla, rd,ti,sg

Where,

Sla is Surface Linguistic Analysis
Rd is Redundancy Detection
Ti is Topic Identification
Sg is Summary Generation

(c) Parts of Speech Tagger

Let ST is the set of text file ST=tx1, tx2,...,txn

Where,

tx1,tx2,txn are the set of text files.

(d) Event Extraction
Let EE is the set of .XMI file EE=xmi1, xmi2,..,xmin
Where,
xmi1, xmi2, xmin are the set of text files.

(e) UML Modeling Tool
Let UT is the generation of ER Diagram UT=erd
Where,
Erd is ER-Diagram.

3.4 Mathematical Model
3.4.1 Keywords Extraction
We need to find out important keywords or terms, for this Let,
S= s1, s2 ...sn,
Where,
S is set of all user input sentences

Keywords (W) is defined as,
W= w1, w2, ...,wk
Where,
wk is k number of keywords,
Each keyword weight is calculated by using
Tfi,j= Σsi,j

Where,
Sij is the occurrence of keyword j in sentences i

If $tf(j) > threshold$
Add in final keyword List

$$KW(S_{i.k}) = \frac{Kywordcount\ \ (S_{i.k})}{lengt\ h\ (S_{i.k})} \ldots\ldots\ldots\ldots\ldots\ldots 1$$

Where
$Kywordcount$ is total no of occurrence of final keyword i in sentence Sk,
Length is Total no of words in sentence Sk,
i is sentence number,
k is document number.(In our system k=1).

Proper Noun Feature:
In general the sentence that contains more proper nouns is an important one and it isMost probably included in the document summary.
Proper nouns (PN) in the sentence is calculated by

$$PN(S_{i.k}) = \frac{PNcount\ \ (S_{i.k})}{lengt\ h\ (S_{i.k})} \ldots\ldots\ldots\ldots\ldots\ldots 2$$

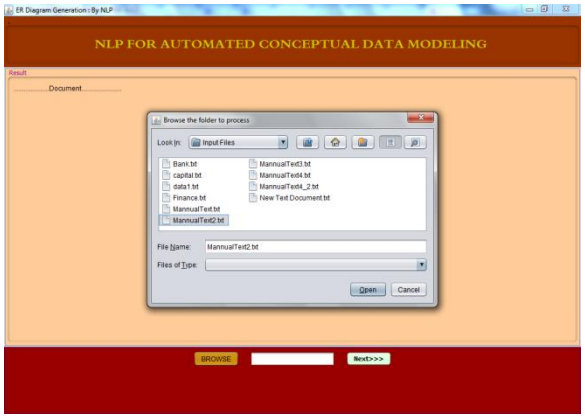
Where
PNCount is total no of nouns in sentence Si:k,
Length is Total no of words in sentence Si;k,
i is sentence number,
k is document number.(In our system k=1).

IV. Experimental Results And Discussion

For Study of "Automatic ER diagramcreation" framework we tested distincttextdocumentsone after another. Proposed framework producesvery good summary of every documents and after that it produces XMI file of every document. Framework uses Star UML modeling tool for perspective ER outline. Every report XMI record is open in this setup and result demonstrates that ER diagram is produced exceptionally well.
At last as per the outcomes acquired by above test on distinct records, we can say that our framework gives client fulfillment, well rundown and programmed create ER diagram.

1. Input file

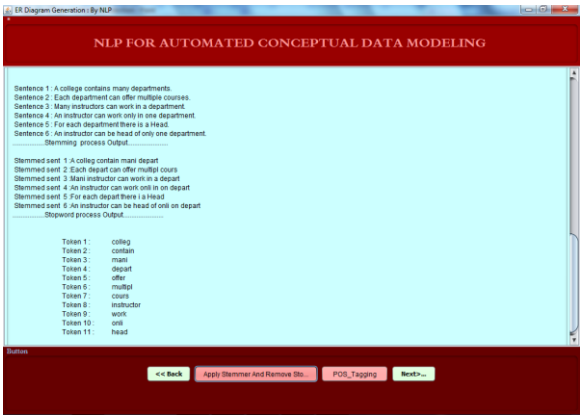
Following screenshot shows the input file which takes the input as paragraph. Here the input file is simple text file.



Screenshot 1: Input file selection

2. Preprocessing(Remove Stemmer and stop words)

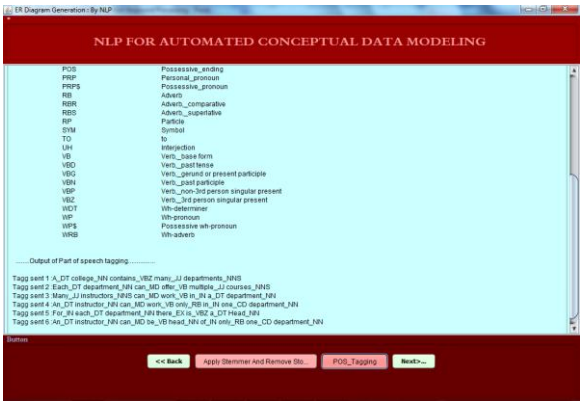
Following screenshot shows the processing in which stemmer is applied and the stop words like "a", "an", "the" etc. are getting removed from the paragraph. Non-significant words are removed from the text. Stemmer performs the process of reducing words to their stems, ie, the portion of a word that is left after removing its prefixes and suffixes.



Screenshot 2: Stemming and Stop word processing

3. Preprocessing(POS Tagging)

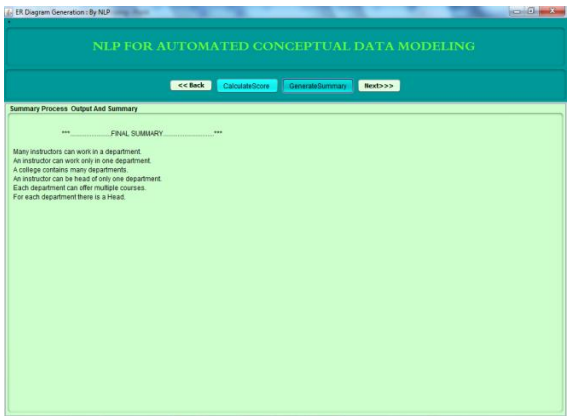
Following screenshot shows preprocessing in which the Part of Speech Tagging to each word,then the words in a sentence are divided and labeled with specific documentations like NNP for Proper person noun, place or thing, NP for noun phrase, VP for verb phrase, and so on.



Screenshot 3: POS Tagging

4. Summary Generation

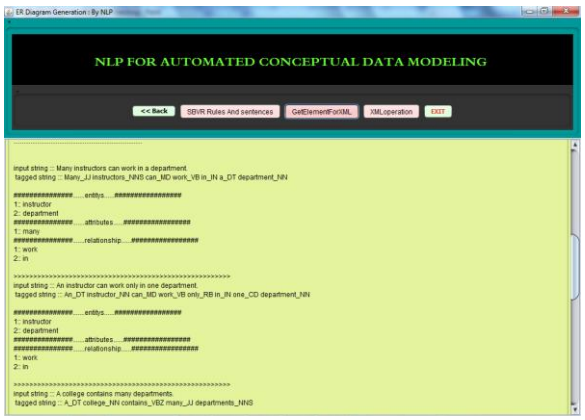
Following screenshot shows final summary generation based on weight of each sentence, this system will then select top weighted sentences as summary of all paragraphs or documents. Here based upon CQP (Code Quantity Principle) we are giving score to each and every sentence, hence the sentences which are having highest score among them, only those sentences will be processed into the paragraph.



Screenshot 4: Final Summary

5. Extraction of Elements

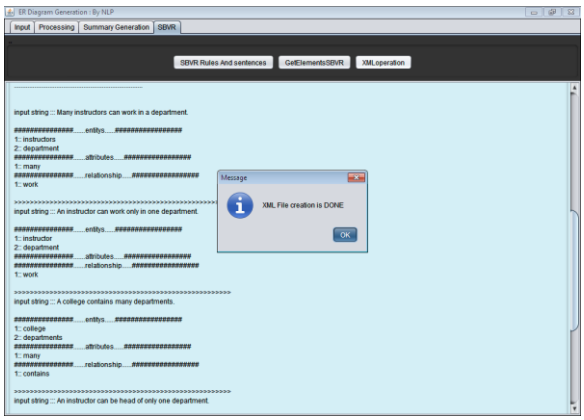
Following screenshot shows the extracted element like nouns, proper nouns, verbs etc. which are very frequently used in the paragraphs. And these elements are the actual elements required to build or draw the E-R diagram. These elements are nothing but the entities and their relationships.



Screenshot 5: Get Elements for ER

6. XMI file creation

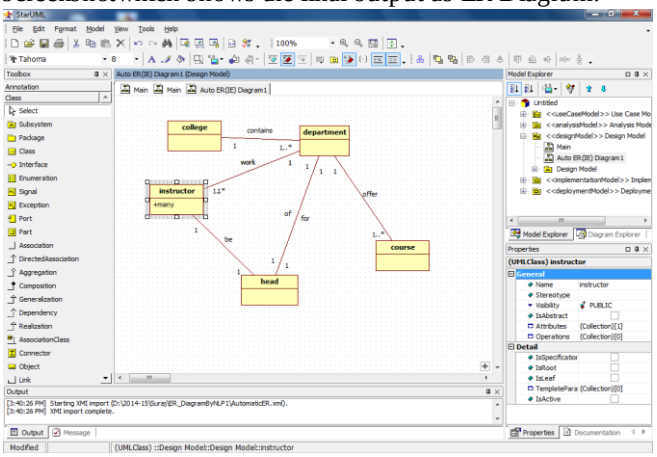
Following screenshot shows the XMI (XML Metadata Interchange) file creation after element Extraction.



Screenshot 6: XML file Generation

7. (E-R DiagramFinal Output)

Following is the screenshotwhich shows the final output as ER Diagram.



Screenshot 7: Final E-R Diagram

Here the created XMI file is imported through the UML modeling tool and the result is displayed.

V. Conclusion

Proposed system is used to create summary of given input datawhich is based on code quantity principle.The input data is nothing but the English paragraphs, these paragraphs are processed and summery is generated. This system termed as "Automated Conceptual Data Modeling" which Generates ER diagram utilizing SBVR terminology on various sentences has been created for extracting the obliged data from the naturallanguageinput which will be utilized to create a XMI document and thus produce an Entity Relationship graph. In future distinct strategies of Entity Extraction can be used for enhancing exactness of ER graph.

For further enhancement of the given system, we can use different modelling tools for generating ER diagram. The other enhancement is to work on sentence ambiguity for better result. This System can be further Enhanced as a Web Application for generating Automatic E-R diagram.

References

[1] Pranitha Reddy, R C Balabantaray, "Improvisation of the DocumentSummarization by combining the IR techniques with Code-Quantity andAttention Linguistic Principles."International Journal of Engineering andInnovative Technology. (IJEIT) Issue 2012.

[2] Elena Lioret, Maria Teresa Roma-Ferri, Manuel Palomar, "Compendium: A text summarization system for generating abstracts of research papers", Data & Knowledge Engineering, Science Direct 88 (2013) 164-175.

[3] Brill, E."A Simple Rule-Based Part of Speech Tagger. In: Proceedingsof the Third Conference on Applied Natural Language Processing", ACL,Trento, Italy (1992) 152-155.

[4] Buchholz, E., Cyriaks, H., Dusterhoft, A., Mehlan, H., and B. Thalheim.,"Applying a Natural Language Dialogue Tool for Designing Databases. In:Proceedings of the First Workshop on Applications of Natural Languageto Databases."(NLDB'95), Versailles, France (1995) 119- 133.

[5] Eick, C. F. and Lockemann, P.C."Acquisition of Terminology KnowledgeUsing Database Design Techniques". Proceedings ACM SIGMODConference, Austin, USA (1985) 84-94.

[6] Gomez, F., Segami, C. and Delaune, C. "A system for the semiautomaticgeneration of E-R models from natural language specifications". Data andKnowledge Engineering 29 (1) (1999) 57-81.

[7] Nazlia Omar, Rosilah Hassan, Haslina Arshad, ShahnorbanunSahran,"automation of Database Design through Semantic Analysis" Proc. Ofthe 7th WSEAS Int. Conf. On computational intelligence, man-machineSyatems and cybernetics (CIMMACS '08).

[8] Imran Bajwa, Mark G. Lee BehzadBordbar, "SBVR Business RulesGeneration from Natural language Specification" in Artificial Intelligencefor Business Agility Papers from the AAAI 2011 Spring Symposium(SS-11-03), pp 2-8, 2011.

[9] Thomas, Pete; Smith, Neil and Waugh, Kevin (2008). Automatic assessment of sequence diagrams. In: 12th International CAA Conference: Research into e-Assessment, 8-9 July 2008, Loughborough University, UK.

[10] P.G. Thomas, N. Smith, K. Waugh "Automatically assessing graphbaseddiagrams" Centre for Research in Computing the Open UniversityWalton Hall Milton Keynes MK7 6AA.