International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Survey on Automated Conceptual Data Modeling

Suraj A. Jogdand[1], Pramod B. Mali[2]

[1]Department of Computer Engineering STES'S, Smt. KashibaiNavale College of Engineering Vadgaon BK, Pune, India

[2]Assistant Professor, Department of Computer Engineering STES'S, Smt. KashibaiNavale College of Engineering Vadgaon , BK, Pune, India

Abstract: ***Database modeling is a difficult task because of its conceptual nature and technicality. With the aim to design databases, numbers of researchers have tried to apply natural language processing in extracting knowledge from requirements specifications. Then again, research on the establishment and utilization of heuristics to help the development of intelligent databases from natural language has been rare. This paper focuses on the issue of extracting semantic knowledge in the creation of ER models from language specifications. The application of semantic heuristics is presented as the method to get the significant ER components, for example, relationships, entities and attributes from the specifications. Prior exploration has demonstrated that syntactic heuristics delivered great outputs about recognizing the important and right results of the ER elements regarding review and accuracy.***

Keywords: Database modeling, natural language processing, ER models, lexical knowledge, semantic heuristics.

1. Introduction

Database modeling can be an overwhelming assignment to both students and designers indistinguishable because of its unique nature and technicality. Much research has been endeavored to apply natural language processing in concentrating knowledge from requirements provision with the mean to design databases. Then again, look into the framing and utilization of heuristics to help the development of intelligent databases from natural language has been rare.In this paper, we are exploring utilization of semantic heuristics in the era of ER models from natural language specifications. The semantic heuristics will be utilized to focus the significant ER components, for example, entities, attributes and relationships from the database specifications. The application of the heuristics would be acknowledged through the augmentation of a created instrument called ER-Converter [8],[9]. Syntactic heuristics [10] have been actualized in ER-converter. ER-Converter has been assessed against a set of database issues and attained 90% review and 85% accuracy. So as to further enhance the exactness of the results, semantic heuristics are proposed.

2. Literature Review

This segment gives a concise synopsis on data modeling which presents the idea of ER Model and surveys the past work that applies natural language processing to Databases, the current system, strategies and constraints are talked about. A portion of the work like Data Model Generator (DMG)[12] gives a premise to the advancement of new heuristics connected in ER-Converter.

2.1 Overview of Data Modeling

The initial phase in designing a database application is to realize what data the database must store. It is called as requirements analysis stage. The data assembled in this step is utilized to create a high-level state description of the data to be put away in the database. This step is alluded to as conceptual design, and it is frequently completed utilizing the ER model.ER models [3] are constructed around the fundamental ideas of entities, attributes, relationships and cardinality. An entity is an object that exists in this present reality and is discernable from different objects. These are regularly gotten from verbs. Cases of entities incorporate the accompanying: a "student ", a "worker" and a "book". A gathering of comparative entities is called an entity set. A set of attributes is utilized to depict an entity. The level of detail at which we wish to represent to data about entities is reflected by using the attributes of an entity. Attributes may be gotten from adjectives and adverbs. Case in point, the "Student" element set may have "Id_number", "Name", "Location", "Course" and "Year" as its attributes. A relationship is a connection among two or more entities. Relationships can be derived from verbs. For instance, we may have a relationship from this sentence: A student may "take" numerous courses. "take" advises a relationship between the substance "student" and "course". Cardinality addresses to the key constraint in a relationship. In the above example, the cardinality is said to be many-to-many, to demonstrate that a student can take number of courses and a course can be taken by number of students. In an ER diagram, an entity is usually represented to by a rectangle. An ellipse generally represents to an attribute and a diamond shape demonstrates a relationship. In cardinalities, 1 is used to represent one-sided and many-sided is represented by M.

2.2 Applying Natural Language Processing (NLP) to Databases

Much work has endeavored to apply natural language in extracting knowledge from requirements specifications or dialog sessions with designers with the intent to design databases.

Dialogue tool [2] is a knowledge based tool connected to the German language for delivering a skeleton diagram of an Enhanced Entity-Relationship (EER) model. This tool is a piece of a bigger database outline framework known as RADD (Rapid Application and Database Development) which comprises of different segments that structure a complex tool. So as to get information from the designer, a directed dialogue is built amid the design process. The

conversion of the structure of natural language sentences into EER model structures is a process which is focused around heuristic suppositions and pragmatic interpretation. The point of the pragmatic interpretation is the mapping of the natural language input onto EER model structures utilizing the outputs of the syntactic and semantic study. One significant limit in this framework is that the precision of the EER model delivered relies on upon the size anddifficulty of the grammar used and the scope of dictionary.

ANNAPURNA [4] is planned to present a computerized environment to semi-automatic database design from knowledge obtaining up to producing an ideal database blueprint for a given database management system. ANNAPURNA focused on the stages concerned with securing the terminological standards. The initial phase in securing of the terminological knowledge includes extracting the knowledge from queries and rules that have the type of natural language statements. The knowledge acquired would then be put into the manifestation of S-Diagrams. An S-diagram is a data model that is utilized to specify classes, subclass associations between classes and attributes. The restriction of the S-diagrams is that S-diagrams perform best when the complexity is less.

DMG [11] is a principle based configuration tool which keeps up principle and heuristics in a few knowledge bases. A parsing algorithm which getsinformation of a grammar and a dictionary is intended to meet the prerequisites of the tool. In the middle of the parsing stage, the sentence is parsed by retrieving vital data from the grammar, presented to by syntactic rules and the dictionary. The parsing results are processed further on by principles and heuristics which set up a relationship in the middle of semantic and design knowledge. The DMG needs to communicate with the user if a word does not exist in the vocabulary or the data of the mapping standards is confusing. The linguistic structures are then converted by heuristics into EER ideas. Despite the fact that DMG proposed a substantial number of heuristics to be utilized as a part of the transformation from natural language to EER models, the tool has not yet been created into a practical framework.

E-R generator [5] is an alternate rule based framework that creates E-R models from natural language specifications. The E-R generator comprises of two sorts of rules: specific rules associated to semantics of a few words in sentences, and generic rules that distinguish entities and relationships on the premise of the logical form of the sentence and on the premise of the entities and relationships under development. The knowledge representation structures are developed by a Natural Language Understander (NLU) framework which utilizes a semantic interpretation approach. There are circumstances in which the framework requires help from the user keeping in mind the end goal to resolution confusions, for example, the connection of attributes and determining anaphoric references.

CM-Builder [7] is a natural language based CASE tool which intends for supporting the analysis phase of software development in an object-oriented framework. The tool utilizes natural language processing systems to break down software requirements documents and produces beginning conceptual models represented in Unified Modeling Language. The framework uses discourse interpretation and frequency analysis in creating the conceptual models. CM-Builder still has some restriction. For instance, connection of post modifiers, for example, prepositional expressions and relative clauses is constrained. Different deficiencies incorporate the condition of the knowledge bases which are static and not effortlessly updateable nor versatile. Heuristics, in view of phonetic rules, are accounted for to be used in huge numbers of the frameworks like ANNAPURNA [4], DMG [11] and RADD [2]. The heuristics displayed, on the other hand, are fundamentally focused around language structure. This exploration plans to fill in the crevice by proposing another set of semantic heuristics.

Eric Brill [1] had introduced a basic rule-based part of speech tagger which executes and in addition existing stochastic taggers, but had important advantages over these taggers.The tagger has a great degree of portability. Number of the higher level procedures utilized to enhance the performance of stochastic taggers would not promptly exchange over to a different tag set or type, and unquestionably would not exchange over to an alternate language. Everything aside from the proper noun searching method is consequently obtained by the rule-based tagger [6], making it significantly more portable than a stochastic tagger. In the event that the taggers were prepared on a different corpus, a different set of patches suitable for that corpus would be discovered consequently. Substantial tables of detail are not required for the rule-based tagger.In a stochastic tagger, countless lines of factual information are required to capture logical information. This information is typically a table of trigram detail, showing for all labels taga, tagb and tagc the probability that tagc takes after taga and tagb. In the rule-based tagger, context oriented information is captured in less than eighty rules. This makes for an a great deal more perspicuous tagger, supporting in better understanding and rearranging further understanding of the tagger. Contextual information is presented in a significantly more reduced and reasonable structure. As can be seen from looking at error rates, this minimized representation of contextual information is generally as powerful as the information covered up in the huge tables of context oriented probabilities.

3. Conclusion

In this paper, we have discussed Automation of ER Modeling through Natural Language Processing. Also we have discussed many existing systems and the development which can be done in this topic. CM-Builder, E-R generator, Dialogue tool, ANNAPURNA, DMG and RADD are few existing system which focuses on Automation of ER Modeling through Natural Language Processing. Each and Every system is having its own advantages as well as drawbacks; we will try to improve the drawbacks by generating our own unique system.

References

[1] Brill, E.: A Simple Rule-Based Part of SpeechTagger. In: Proceedings of the Third Conferenceon Applied

Natural Language Processing, ACL,Trento, Italy (1992) pp. 152-155.

[2] Buchholz, E., Cyriaks, H., Dusterhoft, A.,Mehlan, H., and B. Thalheim.: Applying aNatural Language Dialogue Tool for DesigningDatabases. In: Proceedings of the First Workshopon Applications of Natural Language toDatabases (NLDB'95), Versailles, France (1995)119- 133.

[3] Chen, P.P.: English Sentence Structure andEntity-Relationship Diagram, InformationSciences, Vol.1, No. 1, Elsevier (1983) 127-149.

[4] Eick, C. F. and Lockemann, P.C.: Acquisition ofTerminology Knowledge Using Database DesignTechniques. Proceedings ACM SIGMODConference, Austin, USA (1985) 84-94.

[5] Gomez, F., Segami, C. and Delaune, C.: Asystem for the semiautomatic generation of E-Rmodels from natural language specifications.Data and Knowledge Engineering 29 (1) (1999)57-81.

[6] Green, B. and Rubin, G. Automated Grammatical Tagging of English. Department of Linguistics, Brown University, 1971.

[7] Harmain, H.M. and Gaizauskas, R. CM-Builder:A Natural Language-Based CASE Tool forObject-Oriented Analysis. Automated SoftwareEngineering 10 (2) (2003) 157-181.

[8] Omar, N., Hanna, P. and Mc Kevitt, P.Acquisition of Entity-Relationship Models fromNatural Language Specifications UsingHeuristics, 3rd International Conference on IT andMultimedia, UNITEN, Malaysia (2005) CDROM.

[9] Omar, N. Heuristics-Based Entity RelationshipModelling Through Natural LanguageProcessing. PhD Thesis (2004). University ofUlster, UK.

[10] Storey, V.C. and Goldstein, R.C.: A Methodology for creating user Views in Database Design. ACM Transactions on Database Systems 13 (3) (1988) 305-338.

[11] Tjoa, A.M and Berger, L.: Transformations of Requirements Specifications Expressed inNatural Language into an EER Model.Proceeding of the 12th International Conference on Approach, Airlington, Texas, USA (1993)206-217.