Question answering through unsupervised knowledge acquisition

Rivindu Perera[1], Udayangi Perera[2]

Department of Computer Science, Informatics Institute of Technology
Colombo 06, Sri Lanka

[1]rivindu.perera@hotmail.com
[2]udayangi@iit.ac.lk

Abstract— **Current question answering systems are usually based on a knowledge base which is populated with domain specific knowledge and managed through Unstructured Information Management Architecture (UIMA). But drawback in this approach is that knowledge base may be grown with knowledge which is not relevant to the users connected with the system. In order to address this drawback we propose unsupervised knowledge accumulation algorithm which can monitor user preferences and acquire knowledge without any supervision of the system management unit. Basically, this algorithm learns domain of interest of each and every user connected with the system and extract knowledge from the web or from a given corpus. We have also adopted several Natural Language Processing algorithms to design this high-level algorithm. Knowledge modelling is done through a conceptual graph based knowledge base. This novel paradigm is evaluated with the help of several connected users and with more than 280 questions. We have achieved excellent accuracy during the evaluation phase. It shows our novel approach is effective and can be used to address the drawback decently.**

Keywords— **Knowledge acquiring, question answering, knowledge management, Natural Language Processing, conceptual graphs**

I. INTRODUCTION

Knowledge acquisition is one of the challenging tasks in developing knowledge bases for question answering systems. Therefore, several approaches are used to acquire knowledge which covers vast amount of aspects. However, all these methods work totally separated from the user communities. This simply encompasses that always knowledge engineers working behind the interface must look for knowledge to fill the knowledge base in all domains regardless of what is important for the addressed user community. But when considering the time and the resource consumption there are limitations in this method which cause these knowledge bases to be useless for the connected user community even though they are enriched of knowledge in different categories [1].

To address this drawback, this research suggests a novel algorithm which monitor user preferences when question processing. This approach composes the goal as an evolving strategy to acquire knowledge. But the most important area to analyse here is that the source of the knowledge. This may be in either as a web based text resource or as a corpus. Therefore, solution should be able to cater both areas with one generic methodology. Next it is also important to analyse the filtering technique for the acquired knowledge, which means that if a corpus is given then converting the entire text to knowledge may represent a pathetic approach to carry out.

We incorporate amalgamation of relation extraction with Conceptual Graph (CG) generation process to develop the generalized schema to map information to knowledge regardless of the external representation, web information or corpus based text resource. For the filtering process, we integrate Named Entity Recognition (NER) where only agent of the sentence with identified named entity type is considered as a potential relation to convert to a CG. This elementary process is empowered by several algorithms assigned to process the task with high efficiency and accuracy.

However, so far researches in the automated knowledge acquisition are not carried out with this level of competence [2] [3]. Main reason behind this is that currently researches in the question answering systems are focusing on the language processing algorithms and not on the backend layer enriching technologies. This negligence has caused knowledge bases of such systems to grow rapidly with useless knowledge as such user preference monitoring capability is not present [4].

Our research is therefore focused on addressing drawback we have analysed through the algorithm devised which is named as the unsupervised knowledge accumulation algorithm. Practical effectiveness of this algorithm is analysed through the developed prototype – Scholar [5], question answering system based on the proposed algorithm.

II. BACKGROUND OF THE STUDY

Designed algorithm is based on several existing Natural Language Processing (NLP) algorithms and some novel approaches which differ it from a pure heuristic. In this context we present the contribution from the past researches which helped to develop this algorithm and justifications why certain algorithms are not considered.

A. User profiling

Several strategies are present to identify the domain of interest of a user. But when it is applied to question answering systems based knowledge base development only few of them provides high applicability.

Cufoglu et al. [6] present a Weighted Instance Based Learning approach for user profiling and this methodology is depending on Per Category Feature (PCF) as this approach is an extended version of Instance Based Learning (IBL). Though this approach sounds as a better approach for user profiling when used in a question answering system several drawbacks can be arisen. For an example, due to predefined weighting strategy which will be executed in the pre-processing stage, distance measurement can generate inaccurate results in an environment where natural language is used.

But ontology based user profiling introduced by Pan et al. [7] can be considered as an approach which is more suitable in a computational linguistic background. In this research guided by Pan and his team, introduce the user ontology modelling through the identified instances of terms. Therefore, this tactic is even more user centred. User ontology (Θ) formal

description (1) shows the effectiveness of the method initialization to suit with what is expected by a user profiling approach.

$$\Theta = (C, R, I, A) \quad (1)$$

Where *C* represents the concept set, *R* represents set of relations and *I* symbolizes set of instances. Furthermore, set of axioms which will stand for rules is defined through *A*. As this method captures vast amount of information about the user profiling, it is more practical to be used in a system where preferences of users show high diversity. But drawback noticed in this scheme is that excessive semantic richness can also tend to gather inaccurate preference orders while training. To address this control plan is required which on the other hand will not decrease the effectiveness. Furthermore, research carried out in the domain of evolving user profile creation by Iglesias et al. [8] also supports the idea presented by Pan and his team. But also emphasizes that such evolving methods are easy to be used but difficult to be controlled when data set and number of connected users are increased rapidly over time. Therefore, this provides an indirect suggestion to control the user profiling via the system rather allowing it to be totally unsupervised.

B. Relation extraction

In this context, focus is placed on the relation extraction from a given corpus. Overview of task is to extract agent-verb-patient relations from the text resource considered.

Yang et al. [9] introduce clustering based approach for relation extraction considering the syntactic dependency of verbs in a given context. In this research the main drawback identified is that it requires extensive background information about the thematic role of the verb to extract the relation accurately. But in a question answering domain, especially when considering open domain question answering this type of information supplication is hard to be carried out [10].

However, addressing this limitation two staged semantic relation extraction is first introduced by Fu et al. [11] through an empirical research on the relation extraction domain. Though this method relies on both pattern based and association rule which is supported by clustering approach there is no considerable F-measure value to term this approach better than other strategies available. For an example Maximum Entropy (MaxEnt) based relation extraction discussed by Suxiang et al. [12] in language sensitive environment shows 90%-96% accuracy. But when using MaxEnt, it is required to be used with an exhaustively trained models which may also consume some amount of resource while decoding the relation. This emphasizes that even though the training cost is ignored, decoding must be evaluated with a high focus on the resource allocation. In a question answering system this type of issues must be dealt with the most appropriate solution available as later changes are not easy to be incorporated. However, according to Ratnaparkhi [13] as observed expectation ($E_p f_j$) is based on class-context pairs as defined in observed expectation calculation formula (2).

$$E_p f_j = \sum_{a,b} p(a,b)\ f_j(a,b) \quad (2)$$

Where, $p(a,b)$ represents the Probability of pair (a,b) from training set $\{(a_1, b_1), \ldots, (a_n, b_n)\}$ and $f_j(a,b)$ shows the feature function generated using pair (a, b). Therefore, insertion of a condition to maximize the probability of occurrence can increase the expectation thus providing a better decoding quality.

Nevertheless, when comparing MaxEnt approach with other relation extraction methods such as Ontology driven approach proposed by Bellandi et al. [14] and semantic pattern matching approach proposed by Nie et al. [15], MaxEnt has the ability to automate the process with high accuracy where other methods cannot reach. Therefore, MaxEnt shows the highest applicability for relation extraction task in knowledge acquisition for an open domain question answering system.

C. Relation filtering through NER

Even though relation extraction is accurate and efficient, productivity level of it can be greatly decreased by the accuracy of the filtering technique. To design the algorithm we adopt the NER based manoeuvre. This reflects that NER accuracy will put a major effect on the accuracy of our algorithm. Therefore, extended and critical analysis is required to choose the best algorithm to be integrated with our high-level algorithm.

Currently, the simplest NER algorithm is known as the morpheme based chunking and tagging strategy proposed by Fu [16]. Though it is considered as the simplest, usage flexibility of this approach in a multilingual environment is one factor that influences the research communities to work on this. Additionally, F-measure calculation of this strategy also shows that it has achieved high accuracy rate in three different entity types. But the drawback of this approach is that it needs extensive time and resources to generate rule set to extract entity types. Therefore, field like question answering cannot depend on such fewer amounts of named entity types to perform an unsupervised knowledge acquisition, which means that it inspires us to research on other possible areas.

Todorovic et al. [17] present a methodology for NER based on Hidden Markov Models (HMM). But again if HMM is used in an open domain knowledge acquisition arena, due to high data sparseness, joint probability can cause a negative effect on entity extraction. As a solution for this limitation Conditional Random Fields (CRF) approach for NER proposed by Luo et al. [18] can be incorporated. But as CRF shows low training and inference efficiency again the model availability is in a low stage.

However, as Hui et al. [19] define MaxEnt can be used effectively for NER through extensive models already produced using corpora. Therefore, when considering the requirement for already developed models, MaxEnt shows several good features. Also as a model based on conditional probability, it has shown high accuracy levels as well in several tests carried out by different researches [20] [21] [22].

D. Summary

Extended background research carried out to identify individual algorithms to integrate to our high-level algorithm shows us that user profiling must focus on the term identification strategy and use them to create the preference order for each user rather applying a context based relation to each user. Additionally, classification problems like relation extraction and NER can be solved using MaxEnt approach rather using CRF models though these have slightly increased accuracy when compared to MaxEnt. The reason behind this is that accuracy is not the only criteria that can be considered in designing an algorithm to acquire open domain knowledge to be used with question answering systems.

III. METHOD

We have adopted best set of techniques and algorithms with processes to implement them in a computational background and here we have summarised the development of our new algorithm.

A. Research design

We employed a randomized block design to perform this research. In addition major emphasis is also placed on both

qualitative and quantitative aspects while devising the solution algorithm. Reason behind this selection is that knowledge acquisition is an area where qualitative measurement is required while we employ several machine learning algorithms, it is vital to cover quantitative analysis as well.

B. Unsupervised knowledge acquisition

This novel algorithm designed to acquire knowledge from unstructured text resources amalgamated several existing NLP algorithms with the new design strategy proposed. Furthermore, we developed this algorithm on top of the question answering system – Scholar. Therefore, it was easy to measure the real effect of the algorithm as well. Fig. 1 depicts the control flow of the algorithm with a brief overview.

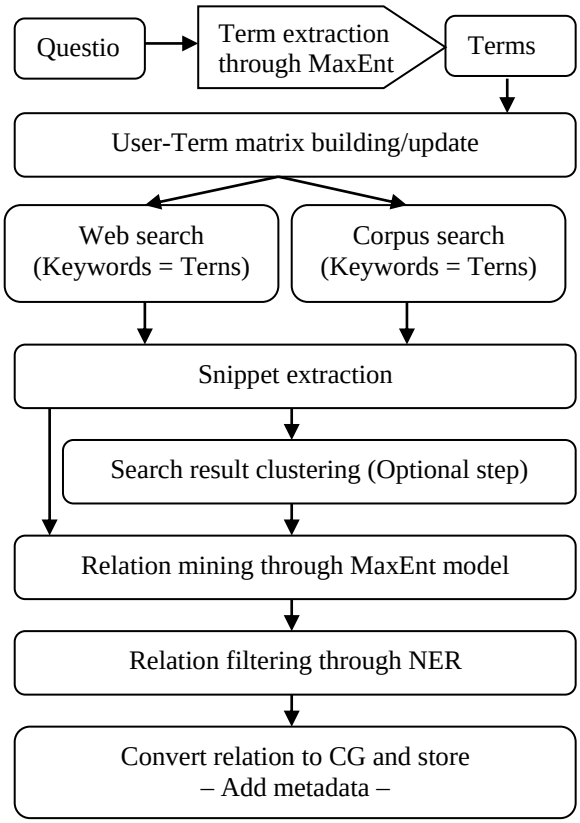

Fig. 1 – Overview of unsupervised knowledge accumulation algorithm

While users are processing questions, terms used in such questions are extracted automatically with the help of MaxEnt model. Also in our first attempts we employed a Bag-of-Word model also with the extracted terms. But in this we have noticed that such term expansion can lead to useless context retrieval and therefore final version of algorithm is only based on pure term extraction.

This acquired list of terms for each user is used to build the user-term matrix which is used to initialize the next few steps. In this level we also use semantic relation mapping within the matrix through automated category identification. Currently, algorithm implementation supports 12 different categories where terms are semantically related. We represent this as a slot based reference formula as shown in (3) expressed through an example.

UTR (u1, utr) ^ User (utr, "Perera") ^ Term (utr, "Computer") ^ Category (utr, "Science") (3)

Where UTR represents user-term reference and utr represent reference instance being used. This extended representation helps to model the semantic relation in a more generic way to use it in later stages.

Each extracted term is used to search the text resource (either web based text or a given corpus) and then snippet is extracted from this search result. At this point as an optional step search result clustering can also be incorporated. But usage of such task can dramatically reduce the amount of knowledge acquired. Therefore, we keep this clustering process as an optional method without placing it in the core knowledge acquisition module.

Relations extracted from the text resources are filtered with the condition that whether agent of the relation is a named entity type or not. Current development of the algorithm supports 28 different named entity types and some of them are shown in Table I below.

Table 1 – SUBSET OF SUPPORTED NAMED ENTITY TYPES

Subset of supported named entity types		
Automobile	Country	Holiday
Anniversary	Drug	Movie
City	Facility	Music Group
Company	Geographic Feature	Natural Disaster
Continent	Health Condition	Operating System

Once the filtering is done for the extracted relations conversion of these relations to CGs is performed via a rule based approach. In this step CG is also enriched with the named entity type assigned for the agent thematic role. Sample CG generated in this phase is shown in Fig. 2 below in the XML format.

```
<cg>
    <connector>provide</connector>
    <agnt etype="Company">Morningstar</agnt>
    <ptnt>Morningstar index data</ptnt>
</cg>
```

Fig. 2 – Acquired relation as a conceptual graph

Implemented question answering system to test the algorithm also supports dynamic visualization of these conceptual graphs in the format defined by Salloum [23]. This visualization is shown in Fig. 3 below.

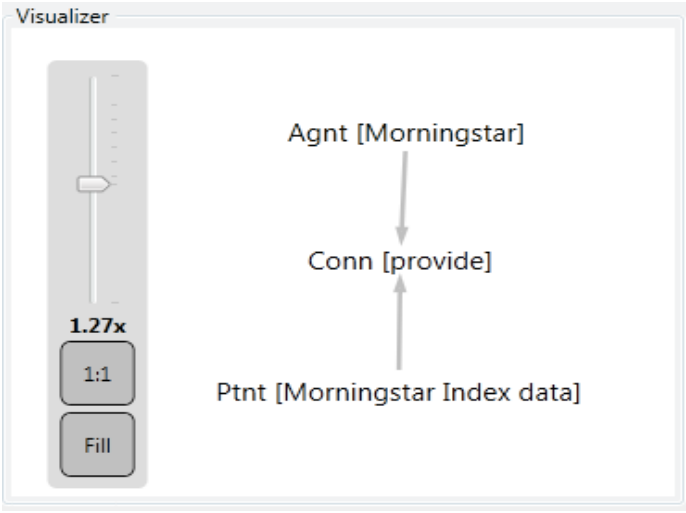

Fig. 3 – Dynamic visualization of acquired relation

This representation of the knowledge supports later extraction of specified knowledge element for a given query through CG projection operations.

C. CG projection

To access the acquired knowledge later in a query processing stage, we incorporate rule based CG projection

process. Two string matching algorithms, Smith Waterman algorithm and Levenshtein distance are mingled to get the required agent of a CG for a given query. But it is also possible to perform reasoning while projection through ontology inspired approach. However, named entity type is also used while performing a CG projection.

IV. RESULTS

Evaluation is carried out configuring the maximum relation extraction limit. Therefore, for each acquired term through user profiling only 20 CGs are generated. Then this acquired CGs are validated to check the erroneous relations extracted and converted to knowledge through CG conversion. For 280 questions executed, system able to find 38 terms associated with questions. Therefore, we noticed that for some questions, no significant term is found which can be used for knowledge acquisition.

During the final evaluation we noticed that system has acquired 760 (38 x 20) CGs. Out of this hefty collection, 163 CGs are considered as invalid due to erroneous results, which includes CGs with erroneous agents, patients or inappropriate connector elements.

Finally, we noticed that system has achieved 78.51% average accuracy value for 38 terms processed. Due to extensive test cases for this evaluation entire result table is difficult to be presented. Set of terms that have achieved more than 50% accuracy level is shown in Table 2 below with their respective valid and invalid CG values.

Table 2 – SET OF TERMS WHICH ACHIVED ABOVE 50% ACCURACY LEVEL

Term	Valid CGs	Invalid CGs	Individual accuracy (%)
Google founder	19	1	95
Basketball	16	4	80
Mt. Kilimanjaro	13	7	65
Nuclear complex	17	3	85
Microsoft	19	1	95
Rochester	14	6	70
Magnetic levitation	18	2	90
Sri Lanka	11	9	55
Hamlet	18	2	90
Niagara	13	7	65
Rome	18	2	90
Share price	12	8	60
Spain	16	4	80
Sacramento	15	5	75

A few worst case scenarios are depicted in Table 3, considering 50% individual accuracy level as an indicator.

Table 3 – ALL SCENARIOS BELOW 50% ACCURACY LEVEL

Term	Valid CGs	Invalid CGs	Individual accuracy (%)
SET	2	18	10
Amazon	9	11	45
NASA	6	14	30
Pitch	1	19	5
Square	5	15	25

In next section, we try to investigate reasons behind this accuracy levels deeply.

V. DISCUSSION AND INTERPRETATION

Result shown in Table 2 shows that unsupervised knowledge acquisition algorithm can acquire considerable amount of knowledge from the web through monitoring terms. But it should also be emphasized that during testing it is noted that relatively uncommon terms (Ex: Sri Lanka) are not competing with other more specific terms in the same category. But when focusing on worst case scenarios, it is clearly identified that acronyms and ambiguous terms have generated more inaccurate results ending up with accuracy levels below 50%.

Term “SET” is an acronym identified as term which stands for Secure Electronic Transfer (SET). Though this is an acronym, as it plays a major role in English as a verb, system has faced a difficulty in extracting knowledge and has ended up with inaccurate connecting elements for CGs generated. On the other hand, reason behind NASA is totally different from the previous scenario analysed. CGs generated for NASA mostly use the denotation as National Aeronautics and Space Administration (NASA). But as in all occurrences that full denotation is associated with acronym, system has considered “Space” as a verb ending up with partial CGs.

For other three terms, “Amazon”, “pitch” and “square”, reason is noticed as term ambiguity. But out of them term which has less ambiguity and which is almost used as a name for an entity has achieved higher accuracy when comparing with other worst cases.

Conversely, we observed that terms which have achieved above the expected 50% accuracy are more specific in terms of the semantic value. Terms like “Google founder”, “Mt.Kilimajaro” and “Niagara” are widely used and precisely described terms in web documents. Due to this reason such terms are mostly associated with more accurate knowledge structures. However, findings of this evaluation phase can be summarised into two major categories. Specific terms which are used in the web documents extensively referring the same resource are associated with more accurate knowledge structures. But acronyms and ambiguous terms mentioned in documents are mostly associated with inaccurate knowledge structures.

These drawbacks noticed in this approach inspire us to research on term weighting factors to be strengthened with more specific criteria, which means that abstract and ambiguous terms should be automatically removed from the user-term matrix to increase the accuracy of the unsupervised knowledge acquisition.

VI. CONCLUSION

This research introduced the novel unsupervised knowledge accumulation algorithm which is developed combining both NLP algorithms and rule based approaches. Implemented question answering system – Scholar is used to test the algorithm by integrating it to the main control flow of Scholar and applying user profiling on top of that. During the evaluation it is noticed that designed algorithm can contribute positively to the knowledge acquisition phase. But implications of the result show that general terms used in user profiling can lead the process to inaccurate results. Therefore, in the future we will focus on enhancing this algorithm with strict term weighting factors to identify most specific terms by reducing abstract phrases. Furthermore, we noticed that it is important to utilize a lexical resource to identify the acronyms more precisely prior to knowledge search.

Therefore, future researches inspired by this excellent research outcome will be focused on algorithms to monitor user profile more precisely providing only important, useful and precise phrases to build the user-term matrix.

REFERENCES

[1] H. Wang, "Knowledge management in small and medium sized enterprises", *Second IEEE International Conference on Information Management and Engineering*, Chengdu. 2010. pp.420-424.

[2] G.B. Winter, "An automated knowledge acquisition system for model-based diagnostics", *IEEE Systems Readiness Technology Conference*, Dayton. 1992.pp.167-172

[3] B. King, A.P. Steward and J.I. Tait, "Towards automated knowledge acquisition for process plant diagnosis", *IEEE Colloquium on Knowledge Discovery in Databases*, London.1995.

[4] G. Sarker, M. Nasipuri and D.K. Basu, "An inductive learning strategy for automated knowledge acquisition based on concept rule", *IEEE Region 10 Conference on Computer and Communication Systems*, 1990.

[5] R. Perera, "Scholar: Cognitive Computing Approach for Question Answering", University of Westminster. 2012.

[6] A. Cufoglu, M. Lohi and C. Everiss, "Weighted Instance Based Learner (WIBL) for User Profiling", *10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics*, Herl'any. 2012.

[7] J. Pan, B. Zhang, S. Wang, G. Wu and D. Wei, "Ontology Based User Profiling in Personalized Information Service Agent", *Seventh International Conference on Computer and Information Technology*, Richmond. 2007.

[8] J. A. Iglesias, P. Angelov, A. Ledezma and A. Sanchis, "Creating evolving user behavior profiles automatically", *IEEE transactions on knowledge and data engineering*, vol.24, pp.854-867, 2012.

[9] Y. Yang, Q. Lu and T. Zhao, "A Clustering Based Approach for Domain Relevant Relation Extraction", *International Conference on Natural Language Processing and Knowledge Engineering*, Beijing. 2008.

[10] C. Wang, A. Kalyanpur, J. Fan, B. K. Boguraev and D.C. Gondek, "Relation extraction and scoring in DeepQA*", IBM Journal of Research and Development*, vol. 56, pp.9:1-9:12, 2012.

[11] J. Fu, X. Fan, J. Mao and X. Liu, "Two Stage Semantic Relation Extraction", *Ninth International Conference on Hybrid Intelligent Systems*, Shenyang. 2009. pp.327-331.

[12] Z. Suxiang, W. Juan, W. Xiaojie and L. Lei, "Automatic Entity Relation Extraction Based on Maximum Entropy", *Sixth International Conference on Intelligent Systems Design and Applications*, Jinan. 2006. pp.1-5.

[13] A Ratnaparkhi, Learning to parse natural language with maximum entropy models. *Machine Learning*. Vol. 34, pp.151-175. 1999.

[14] A. Bellandi, S. Nasoni, A. Tommasi and C. Zavattari, "Ontology-Driven Relation Extraction by Pattern Discovery", *Second International Conference on Information, Process, and Knowledge Management*, St. Maarten. 2010. pp.1-6.

[15] T. Nie, D. Shen, Y. Kou, G. Yu and D. Yue, "An Entity Relation Extraction Model based on Semantic Pattern Matching", *Eighth Web Information Systems and Applications Conference*, Huhehot. 2011. pp.7-12.

[16] G. Fu, "Chinese Named Entity Recognition using a Morpheme-based Chunking Tagger", *International Conference on Asian Languages Processing*, Harbin.2009. pp.289-292.

[17] B.T. Todorovic, S.R. Rancic, I.M. Markovic, E.H. Mulalic, V.M. Ilic, "Named Entity Recognition and Classification using Context Hidden Markov Model", *Ninth symposium on neural network applications in electrical engineering*, Serbia. 2008. pp.1-4.

[18] F. Luo, H. Xiao and W. Chang, "Product Named Entity Recognition Using Conditional Random Fields", *Fourth International Conference on Business Intelligence and Financial Engineering*, Jeju Island. 2011.

[19] N. Hui, Y. Hua, T. Ya-zhou and W. Hao, "A Method of Chinese Named Entity Recognition Based on Maximum Entropy Model", *International Conference on Mechatronics and Automation*, Changchun.2009. pp.2472-2477

[20] W. Jiang, Y. Guan and X. Wang, "Improving feature extraction in named entity recognition based on maximum entropy model", *Fifth International Conference on Machine Learning and Cybernetics*, Dalian. 2006. pp.2630-2635.

[21] S. Biswas, S. Mohanty and S.P. Mishra, "A Hybrid Oriya Named Entity Recognition system: Integrating HMM with MaxEnt", *Second International Conference on Emerging Trends in Engineering and Technology*, Nagpur. 2009. pp.639-649.

[22] A. Ekbal, S. Saha and M. Hasanuzzaman, "Multi-objective Approach for Feature Selection in Maximum Entropy based Named Entity Recognition", *International Conference on Tools with Artificial Intelligence*, Arras. 2010. pp. 323-326.

[23] W. Salloum, A question answering system based on conceptual graph formalism. *Second International Symposium on knowledge acquisition and modeling*. Wuhan. 2009. pp.383-386.