Towards a Question Answering System Based on Precisiated Natural Language

Mehdi Rohaninezhad
School of computer science, Faculty of Technology and Information Science, National University of Malaysia,
mr@ftsm.ukm.my
rohaninezhad@gmail.com

Nazlia Omar
School of computer science, Faculty of Technology and Information Science, National University of Malaysia
no@ftsm.ukm.my

Abstract—precisiated natural language is a new paradigm in soft computing based on information granulation. This paradigm is placed in central position of many investigations in the field of human machine interaction in these years. Precisiated natural language addresses a model for dealing with vagueness of information and suggests a kind of knowledge representation, deduction and computational approach for natural language manipulation. Moreover, ontological architecture is an accepted model for developing this theory. In this work we are going to develop an ontology based question answering system which is one of crucial application of this paradigm.

Keywords-component: precisiated natural language; Computing with words; question answering systems; ontology; fuzzy logic; semantic web

I. INTRODUCTION

Precisiated natural language (PNL) is a new soft computing paradigm which is offered by Zadeh[1] in 2005. Essential supremacy of this paradigm in comparison with classical approach of natural language processing (NLP) is its power in modeling vague information and doing reasoning and deduction in uncertainty situation. As it is common in every soft computing approach knowledge representation is the prerequisite of doing computation and reasoning. Formal knowledge representation in PNL is “X is r R” which is called generalized constraint (GC). X, R and is r are constraint variable, constraint relation and constraint index respectively. These are different forms of GC,

\[
\begin{align*}
X \text{ is}_d R & \quad \text{Disjunctive possibility} \\
X \text{ is}_p R & \quad \text{Probability} \\
X \text{ is}_v R & \quad \text{Veristic} \\
X \text{ is}_u R & \quad \text{Usuality} \\
X \text{ is}_{fg} R & \quad \text{Fuzzy graph} \\
X \text{ is}_{rs} R & \quad \text{Fuzzy random set} \\
X \text{ is}_g R & \quad \text{Group}
\end{align*}
\]

in which probability, possibility and veristic are the main form of GC and other forms can be generated via these three main propositions.

Actually every precisiable natural language proposition should be translated to this form before any computation and reasoning. On the other hand, ontological approach as the explanatory database is accepted model for translating natural language proposition to canonical form of PNL. Due to this reason we have suggested ontological model for our question answering system (QAs). The main subjective of this paper is offering PNL-Based framework for restricted domain QAs. Open domain QAs such as web search engines entail more investigation in the field of common sense reasoning [2] and implies more complexity. It should be noted that in this work we are dealing with subjective question rather than objective[3].

II. EXPLANATORY DATABASE

A. Ontological Model

In this paper ontology is viewed as the explanatory database. More precisely, ontologies play the role of developing concepts and relations for QAs. Some of challenging problem of PNL-Based QAs are finding relevant facts to the input query, classification between PNL and non PNL facts and defining metric of concepts. Explanatory database is viewed as a foundation to give a solution for these difficulties. Ontology web language (OWL) which is standard of W3C [4] for semantic web is assumed as the platform for developing PNL-Based QAs in this work. OWL description logic is a kind of formalism knowledge representation based on some axiom such as type definition restriction, however, it has enough power to model information and prepare infrastructure for reasoning and computation. In this investigation architecture of ontology is based on two main categories of ontologies, domain ontologies which is consisted of ontology of person, medicine, illness and etc and GC ontology that plays the role of precisiation of vague information for domain ontologies. Actually the architecture of ontology in this work is a continue of Reformat et.al[5]. However this work has offered a novel approach for modeling inter granular proposition[6] in constraint ontology.

In the case of GC ontology X, is, and R form the triple of subject, predicate and object respectively according to OWL DL standard. As a simple example, assume the proposition “John is young”, thus GC form of this sentence is symbolized like “Age (John) is$_d$ young”, so the model of its GC ontology is showed Fig.1.

![Fig.1. OWL model of sample GC “Age(John) is$_d$ young”](attachment:owl-model.png)
Constraint ontology is rooted on fuzzy concepts hierarchy like fuzzy membership function (MF), \( \alpha \)-cut, terms, modifier and quantifiers[7-9]. As it is represented in Fig.2, all of constraint ontology axioms have the \( ED:is \) prefix which stands for explanatory database. According to this figure \( ED:is \) is the object property that connects \( ED:ConstraintVariable \) to \( ED:ConstraintRelation \). It should be noted that in PNL paradigm each constraint variable may be a fuzzy event, function of other variable, proportion of other variable, groups of variables or etc[10]. More specifically many of natural language propositions are inter granule. For instance, let us assume \( p \) as proposition, thus,
\[
p: \text{it is true that John likely is young}
\]
\[
p^*: [\text{[Age(John) is \_young is \_true]}] \text{ is true}
\]
\[
p^{**}: [\text{[A(B) is \_C is \_D]}] \text{ is, E}
\]
in which \( p \) is the proposition, \( p^* \) is its translation to GC and finally \( p^{**} \) is its protoform[11]. As it is shown in this example each constraint variable itself is another GC proposition, so constraint variables are divided to two main categories in this work \( ED:N\_Array \) and \( ED:FuzzyEvent \). That \( ED:N\_Array \) represent the n-array constraint variable and \( ED:FuzzyEvent \) shows fuzzy events.

Fig.3 shows the ontology of constraint relations in which \( ED:ConstraintRelation \) class is pointed to \( ED:FuzzyModifier \) and \( ED:FuzzyMF \) classes via \( ED:f\_modifier \) and \( ED:f\_mf \) object properties respectively. \( ED:FuzzyMF \) plays the role of developing the fuzzy MF of each constraint relation. According to Matlab fuzzy toolbox there are several different shape of membership function like PI, S, Z, Bell or etc[12]. And each of these membership functions can be drawn via 3 or 4 fuzzy pair. So each MF figure is connected to three or four \( ED:FuzzyPair \) by \( ED:f\_param \). Each fuzzy pair has the two datatype properties, degree and value that represent membership degree (\( \mu \)) and value of fuzzy object respectively. For illustration, let us assume \( \text{young} \)’s membership function as a trapezoidal function, so its shape can be drawn via four fuzzy point \([0,0,30,45]\) in Matlab toolbox. Now we can model the simple proposition “John is young”, in our ontology as Fig.4.

Fuzzy modifiers like very, not, quite are modeled via \( ED:FuzzyModifier \) class that play the role of constraint relation modification. For instance, assume \( p \) is “John is very young” thus, if \( \mu \) represents membership function of \( \text{young} \) then \( \mu^2 \) will be the membership function of very young. If \( \mu \) is the membership function of a fuzzy term then modified membership function of very, not and quite are \( \mu^2, 1- \mu, \sqrt{\mu} \) respectively.

III. ARCHITECTURE OF QUESTION ANSWERING SYSTEM

Question answering systems are placed in central place of PNL paradigm. Developing question answering systems is viewed as the main soft computing challenging for human machine interaction. PNL claims that classical approaches of natural language processing which mainly are based on classical statistic are not adequate for modeling natural language. Actually many of human concepts essentially are fuzzy and dealing with them via crisp methodologies is not coinentious with those definition and boundaries[13]. Moreover, this theory addresses a methodology for doing computation, reasoning and knowledge representation which can solve some of QAs problems. For instance finding relevant fact to the input query is one of challenging problem of QAs which can be solved by computational approaches of this theory[14]. We have focused on restricted domain QAs in this work. And also this system is designed to answer objective question rather than subjective or normative. Our case study in this paper is breast cancer diagnosis. The final goal of the system is estimating the risk of breast cancer (BC) for a woman.
A. Facts representation and Classification

Facts may be classified based on two views. Firstly fuzzy and non-fuzzy facts, that this kind of specification is rooted in classification of concepts. Some concepts essentially are fuzzy like relevant, likelihood, similarity or etc which there is not any conventional measure system for them. However, some other facts are measurable and meanwhile we can also express their value via fuzzy linguistic variable. For instance age is a concept that is measurable and its value can be expressed via a precise number, however, it can be granular by young, middle or old.

However facts can be classified in other aspect. They can be a matter of classification based on their role in QAs. Essentially three kinds of facts can be defined in this system, some facts are about illness and some others are related to the patient and moreover system needs to know some other information about world knowledge for doing inference and deduction. For each kind of facts we firstly introduce them in a simple form of natural language proposition then translate precisiable facts[11] to GC and finally develop them through OWL inference tools. As it will come all of information about the case study “Rose” and BC risk factors are expressed imprecisely, but are precisiable by translating to GC. It is assumed that this phase is done by human assistant and results will be entered to ontology as the system explanatory database.

According to U.S National Cancer Institute (NCI) [15] risk factors of breast cancer are age, pregnancy age, overweight, alcohol drinking and genetics. Based on GC model these concepts are placed in constraint variable position and will be connected to fuzzy linguistic variable (or constraint relation) via object properties (or constraint indexes). For example Fig.5 shows the ontology of drink alcohol habit. According to NCI report person’s age is the main risk factor of BC and the probability of BC for American’s women is estimated as following.

- from age 30 till 39 probability is around 0.43 percent
- from age 40 till 49 probability is around 1.45 percent
- from age 50 till 59 probability is around 2.38 percent
- from age 60 till 69 probability is around 3.45 percent

So this statistic implies that old age increase the risk of BC significantly. Consequently, we can assume that BC risk is a fuzzy function of age and has more impact on BC risk rather than the other factors. Fuzzy functionality is symbolized via fuzzy graph constraint(isfg)[1]. Our fuzzy relation is a Cartesian granule of age and probability of BC (BCP) which is symbolized as following.

\[ X \text{ isfg } R \]

in which \( R = \text{Age} \times \text{BCP} \) and \( \times \) is Cartesian product and can be defined through several fuzzy if-then rules,

If Age is \( X_1 \) then BCP is \( Y_1 \)
If Age is \( X_2 \) then BCP is \( Y_2 \)
If Age is \( X_3 \) then BCP is \( Y_3 \)

\[ \ldots \]

However for simplicity we have treated with age factor like other risk factor and symbolized it through possibility constraint (isd). Consequently the risk of BC can be assumed as the combination of age factor with other risk factors. Following rules are a part of fuzzy if-then facts for BC risk in the system, in which drink habit, overweight, and physical activity are age dependant facts,

- If age is old then age risk factor is high
- If age is old and drink alcohol habit is regularly then alcohol factor is high.
- If age is old and overweight is high then overweight factor is high
- If age is old and physical activity is rarely then physical activity factor is high
- If pregnancy age is young then pregnancy factor is high

![Fig.5: concept of “drink habit” and its relations in constraint ontology.](image-url)
It should be noted that other fuzzy rules can be added to this list for completion. For instance other rules for middle age can be added. Finally the average risk of BC will be calculated through average of risk factors, but it should be mentioned that young pregnancy age factor decrease the risk of BC.

Like every soft computing approach the first phase of doing deduction and computation is translating propositions to a kind of knowledge representation. So the fuzzy rules are translated to GC as following respectively,

- IF Age(X) is old THEN Age_Factor(X) is high
- IF Age(X) is old AND DrinkAlcohol_Habit(X) is regularly THEN Alcohol_Factor(X) is high
- IF Age(X) is old AND Overweight(X) is high THEN Overweight_Factor is high
- IF Age(X) is old AND PhysicalActivity(X) is rarely THEN PhysicalActivity_Factor(X) is high
- IF Pregnancy_Age(X) is young THEN Pregnancy_Factor(X) is high

System also is dealing with other kinds of facts which are about patient whose name is Rose. Let us assume following scenario for Rose,

*She usually drinks alcohol. And she has given her first child two years after her bachelor graduation. She is old, her daily calorie intake usually is high and don’t have normal physical activities.*

GC form of the facts can be formulated as following,

- Drink_Habit(Rose) is usually
- Age(Rose) is old
- [Daily_Calorie_intake(Rose) is high] is usually
- Physical_Activity(Rose) is rarely

It should be noted that the concept of “usually” in first and third propositions are quite different. More precisely, “usually” in first proposition is a kind of repetition constraint so implies possibility distribution of repetition which can be granular by literature like often, usually, sometimes and rarely. But in third proposition it is related to the concept of expected value of a fuzzy event [16]. So it means that Probability of “high calorie intake by Rose” is usually, thus it is a kind of fuzzy probability and its appropriate proposition is intergranular.

Third kind of facts that system is dealing with are ontological world knowledge that system needs to know for making inference and relation between other types of facts. In this case study system needs following facts,

*System should know over calorie intake with low physical activities cause overweight and bachelor graduation is around 24.*

So we have following GC facts:

- (Age(Bachelor_Graduation(X)) is around 24) is usually
- (Daily_Calorie_intake(X) is high) and (Physical_Activity is rarely) then overweight(X) is high

Semantic web rule language (SWRL) is a sub language of OWL and Rule Markup Language. With this additional plugin in Protegé we can make relation between antecedences and consequences[4], SWRL can do deduction, reasoning and generating new OWL axioms based on first order logic via Jess which is its reasoner. OWL sugar fuzzy if-then rules are symbolized as following for age, drink habit, overweight, physical activity and pregnancy age respectively.

- BC:AgePerson(?x) ∧ BC:RiskFactors(?y) ∧ ED:is_d(?x, ED:Old) → ED:is_d(?y, ED:High)
- BC:AgePerson(?x) ∧ BC:DrinkHabit(?y) ∧ BC:AlcoholFactor(?z) ∧ ED:is_d(?x, ED:Old) ∧ ED:is_d(?y, ED:Often) → ED:is_d(?z, ED:High)
- BC:AgePerson(?x) ∧ BC:OverWeight(?y) ∧ BC:OverweightFactor(?z) ∧ ED:is_d(?x, ED:Old) ∧ ED:is_d(?y, ED:High) → ED:is_d(?z, ED:High)
- BC:AgePerson(?x) ∧ BC:PhysicalActivity(?y) ∧ BC:PhysicalActivityFactor(?z) ∧ ED:is_d(?x, ED:Old) ∧ ED:is_d(?y, ED:Rarely) → ED:is_d(?z, ED:High)
- BC:AgePregnancy(?x) ∧ BC:OverWeight (?y) ∧ BC:AgePregnancyFactor(?z) → ED:is_d(?x, ED:Young) ∧ ED:is_d(?y, ED:Young) ∧ ED:is_d(?z, ED:High)

These are just several samples of our system rules and obviously they can be extended by other if-then rules. Actually defining more rules for system prepares a better responsibility, however, entails more complexity.

It should be mentioned that concepts synonyms are developed through annotation properties in this explanatory database. More concretely, synonyms of each concept are expressed in the form of annotation property. For instance concept of approximate can be expressed through different synonyms in a proposition like about, around or almost.

**B. Doing deduction and computation**

Essential supremacy of PNL paradigm is its ability of doing computation and reasoning based on GC knowledge representation[11]. For this purpose fuzzy arithmetic and logic is required as an addition to crisp arithmetic and bivalent logic. Current tools of reasoning for ontological knowledge base are based on crisp arithmetic and logic, however, their trend is going to develop fuzzy arithmetic as Stoilos, G. et.al in [17] have offered. For this aim we have a unit to do fuzzy computation and reasoning. This unit compute the test-score[8, 18] of each GC proposition for giving the degree of meaningful for each GC proposition. More importantly this part of QAs does inference with computational part of PNL[1]. Extension principle, intersection syllogism, probability rule and computational rule of inference are a part of computational rule
of this paradigm. Extension principle is viewed as the main fuzzy arithmetic and deduction rule.

Breast cancer risk is a function of risk factors that are represented in constraint ontology. Now the computational part of the work is estimating MF of probability BC risk which is done based on extension principle. This rule is defined[1],

\[ Y = f(X_1, \ldots, X_n) \]
\[ X_i \text{ is } A_i \]
\[ \mu_B() = \sup(\mu_{A_1}(u_1) \wedge \ldots \wedge \mu_{A_n}(u_n)) \]

Consequently in the case of BC we have following formula,

BC_Risk=f (Age, Pregnancy_Age, Drink_Alcohol, Physical_Activity, Overweight)

Thus,

\[ \mu_{BC\_Risk}(v) = \sup(\mu_{Age}(u_1)^\wedge \mu_{Pregnancy\_Age}(u_2)^\wedge \mu_{Drink\_Alcohol}(u_3)^\wedge \mu_{Physical\_Activity}(u_4)^\wedge \mu_{Overweight}(u_5)) \]

in which MF of each risk factor is defined in constraint ontology. Fuzzy-graph interpolation rule is one of other useful rules in this work. It is defined[1],

\[ \sum_i \text{If } X_i \text{ is A}_i \text{ Then } Y \text{ is B}_i \]
\[ m_i = \sup_{A_i(v_i)}(\mu_{A_i}(v_i)) \]

For instance, mentioned fuzzy relations between age and BCP as if-then rules and a simple fact "Rose is old age" can be deducted that BCP is high. However the \( \mu \) degree of the conclusion should be calculated via mentioned equation.

**CONCLUSION**

This work is tried to offer an ontological approach for PNL-based question answering systems. Ontology in this research is assumed as an explanatory database and is developed based on canonical form of PNL. Different types of facts are implemented in SWRL or OWL syntactic sugar and finally computational part of PNL has applied to achieve the answer of the input query. However, extending GC ontology for developing probability, fuzzy cardinality and other types of GC steel is challenging. Moreover developing PNL based system needs extending fuzzy computational languages in OWL. In spite of these facts it seems that ontological approach is the best model for developing PNL-based systems and architecture of these kinds of systems should be a matter of further researches.

**ACKNOWLEDGMENT**

I owe my deepest gratitude to Prof.Abbasbandy who taught me fuzzy logic and kindly dedicated his valuable time to me whenever I had a question.

**REFERENCES**


