Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1449-1468
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/IJNAA.2021.5791

Designing a persian question answering system based on rhetorical structure theory

Ali Ehsani[a], Seyed Abdollah Amin Mousavi[b,*], Mahmood Alborzi[c], Maryam Rastgarpour[d]

[a] *Ph.D. Candidate, Department of Information Technology Management, Faculty of Management, Science and Research Branch, Islamic Azad University, Tehran, Iran.*
[b] *Assistant Professor, Faculty of Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran*
[c] *Associate Professor,Department of Information Technology Management, Faculty of Management, Science and Research Branch, Islamic Azad University, Tehran, Iran.*
[d] *Assistant Professor, Department of Computer Engineering, Faculty of Computer, Saveh Branch, Islamic Azad University, Tehran, Iran.*

(Communicated by Ehsan Kozegar)

Abstract

Background and Objectives: A question answering system answers questions using natural language processing, a database, or a document set and returns an accurate answer to the user's question. A large number of efforts have been made to design some systems to answer the user's question. However, limited studies have been conducted on the Persian language to extract the answer to the questions with subjects "why" or "how". The scarcity of such studies is attributed to the complexity and time-consuming analysis and processing of the text structure when going beyond the boundaries of a sentence.
Methods: The present study's primary purpose was to analyze Persian text to create a set of linguistic patterns that can perform related information of causal/explanatory text sentences in a general domain. Information retrieval and text structure recognition algorithms were used for data and text analysis, called Rhetorical structure theory. In addition, 70 questions for "why" and 20 questions for "how" were determined for evaluating the system performance, respectively. Finally, the .NET programming language and relational database, and Persian language interpreters were used to design the software system.
Results: Eventually, a system was designed and published to answer the question with subjects

*Corresponding author
Email addresses: ali.ehsani@srbiau.ac.ir (Ali Ehsani), a.mousavi@iauctb.ac.ir (Seyed Abdollah Amin Mousavi), mahmood alborzi@yahoo.com (Mahmood Alborzi), m.rastgarpour@iau-saveh.com (Maryam Rastgarpour)

Received: July 2021 *Accepted:* October 2021

"why" or "how" with general Data Domain.
Conclusion: The system answered 61 questions with a recall rate of 68%. About 55% of the items were correctly responded to according to the signs of inter-sentence relation, while the correct answers to 13% of questions were related to rhetorical relation among the sentences.

Keywords: Question Answering System, Natural Language Processing, Text mining, causal/explanatory relation, Rhetorical structure theory.
2010 MSC: Primary xxxxx (mandatory); Secondary xxxxx, xxxxx (optionally)

1. Introduction

Nowadays, many efforts have been performed to design systems to factoid answer the user's queries. In Natural Language Processing (NLP), Question answering (QA) systems can be developed to general and private domains.

The QA systems are used in several systems, such as Decision Support systems [19], Business Intelligence [4], Interactive systems with a robot-based interface to allow a conversation to imitate human dialogue [18], community QA systems [5], and QA systems in biomedical medicine field [14]. In the Persian language, most of these systems have focused on questions with factoid answers, which can answer these questions with relatively little linguistic knowledge [13].

Analysis of and processing a text structure are very complex and time-consuming when going beyond the boundaries of a sentence. The main novelty of the present study was to analyze Persian text to create a set of linguistic patterns that can perform related information of causal/explanatory text sentences in a general domain to answer the question with subjects "why" or "how".

In the present study, two analytical models were provided to develop a QA system. The first model was pattern recognition, which targets a set of existed linguistic patterns among sentences. This model improved two independent algorithms for discovering the explanatory/causal role regulated by letters. The other model was Text Parser in Rhetorical structure theory (RST) framework. This model was developed based on the strong reliance on the relation between sentences to take some distance of their components from the sentencing range [20]. The text Parser model was made before generating output by pattern recognizer and providing the heuristic rules for constructing the most appropriate structure, including the full text. These models were combined to develop Persian QAS to meet the "why" and "how" questions.

The remainder of this paper is organized as follows. The studies related to QA were examined in next Section. In Section 3, the algorithms of pattern recognition and proof letters were described. In Section 4, system design and implementation were discussed, and the design process was provided. In Section 5, system evaluation was investigated. Finally, in section 6, the results were concluded.

2. LITERATURE REVIEW

A question answering system (QAS) can be classified into two groups according to the type of items to be responded. These groups are Non-Factoid Question Answering (NFQA) and Fact-Based Question Answering (FBQA) systems. The FBQA consists of infinitive syntactic problems without syntax and the expected answer compared to NFQA. Causal and explanatory are the essential questions in this group, where 'how' and "why" are focused. For instance, "how to activate Windows automatic search command in windows?" or "why ice floats on water?" To answer these questions, we need to search argument relations in text, like cause, motivation, and purpose. Therefore, the conducted research about the FBQA system is more than the NFQA system due to the required linguistic knowledge to address such questions.

A large body of research concluded that the rational exploitation of linguistic structure could improve the effectiveness of answers extraction to NFQA, i.e., understand each sentence's role in text and relationship with each other [3]. Therefore, some research explores the semantic relation, focusing on causal relation extraction and detection.

Initial attempts in this regard were to use particular databases and manual codes for causal detection in written text. For instance, a model was constructed in the COATIS system as causal knowledge acquisition between the two French texts. This model was constructed by classifying index verbs manually in the specialized field, with a precision of 85%.

Khoo detected the English linguistic patterns to extract expressed patterns in sentences from medical abstracts explicitly [9]. A text parser, which was developed to convert causality patterns and sentences into conceptual diagrams, indicated the syntactic structure of purpose.

The pattern graphs were matched with sentence graphs to find causal relation and fill the cause-and-effect pattern with the matched text segments to each slot. The 0.41 and 0.48 precision values were obtained to extract cause and effect, respectively.

Moldovan proposed a semi-automatic approach to detect causal relations and use lexical-syntactic patterns [11]. This model was named semi-automatic because of the automatic extraction of the patterns and manually performing the ranking and pattern confirmation. This study focused on the ¡NP1 verb NP2¿ pattern, which indicated that it is the most common internal pattern to show causality. The vocabulary network was used as the primary source of knowledge extracted from the pair of noun phrases. A list of verb phrases was created by searching in several document sets to each extracted pair from the vocabulary network. Eventually, the semantic constraints were imposed on NP2, NP1, and verb to rank patterns and confirm the related verbs of text. The limitations of this model consist of deriving observations statically from the vocabulary network. The test was performed by text set (TREC-9 2000). Two people examined the casualty relation of recursive relation, and the average precision was 65.5%.

Some researchers have used Machine Learning techniques to automatically exploring causal patterns. For example, Blancol provided a model according to [rel C, VP], [VP rel C] in the classification of patterns [1]. This model consists of above half the causal factors of the TREC5 corpus. Here, C is causality, VP is a verb phrase, and rel is relator such as prepositions or conjunctions that are limited to a words event (e.g., as, since, after, and because). We trained an algorithm to learn whether the causal model is taken using an extracted set of syntactic, semantic, and lexical features from the semantic network to pattern validation. In the test phase, the recall and precision for casual cases were 0.84 and 0.95, while they were 0.86 and 0.96 for non-casual cases, respectively. However, this model only can classify the elements by the relation between "because" and "since" correctly.

Sarrouti designed a QA system, called SemiBioNLQA, compared to advanced biomedical QA systems [14]. This system, which operated entirely automatically, could answer a large number of questions. SemBioNLQA returned users' information requirements with returns factoid answers (like "no", "yes", and a particular name) and ideal ones (like a brief paragraph of related data).

Based on the experimental evaluation of biomedical questions and BioASQ answers (2015, 2016, and 2017), SemBioNLQA outperformed the newest advanced systems.

Presently, the community Question Answering (cQA) system is another QA system that can answer as fact and non-fact. Social network sites created communities with different purposes to facilitate the interaction of people. Particularly, cQAs websites are the frameworks with users interested to express information requirements in the form of questions to other users, which can answer these questions. Some of these websites are Yahoo!answers and Stack Exchange. Herrera provided a model that indicated that Twitter conversations could automatically present QA results (4). The importance of different features with more relation in QA ranking was identified in this model. In addition, this method allowed for returned the complex answers to non-factoid questions. Using a hybrid question categorization technique, Sherkat and Farhoodi defined their personal taxonomy and collected their database, including 9,500 questions asking two Quranic experts. The overall precision for coarse-grain categories was 80.5% [15].

Similarly, Razzaghnoori et al. proposed a technique for converting each question to a matrix where each row indicates the word2vec representation of a word. Then, they applied an extended short-term memory network for categorization, which resulted in an average accuracy of 81.77% for three question datasets. Moreover, they presented the University Of Tehran Question Dataset 2016 collected from some kinds of jeopardy games hosted by the official television of Iran [13].

In another study, Boreshban et al. proposed a Persian QA architecture on non-factoid questions that consisted of question processing, document retrieval, and re-ranking. In question processing, component-required preprocessing was conducted, followed by identifying the question topic. An open-source search engine was used in the document retrieval component, followed by applying the machine-learning methods for the re-ranking task. The experimental results represented that the proposed system has achieved 81.29% accuracy and 71.88% mean reciprocal rank. Thus, our proposed system is the first QA system with features in Persian, especially in the religious domain [2].

Likewise, Fakour and Veisi designed and implemented a medical QA system in the Persian language by collecting and structuring a dataset of diseases and drugs. The proposed system included question processing, document retrieval, and answer extraction. For the question-processing module, a sequential architecture was designed that retrieves the central concept of a question by various components. This tool was based on rule-based methods, natural language processing, and dictionary-based techniques. Different customized language processing tools (including part of speech tagger and lemmatizer) were also developed for Persian during this research. Their results showed that this system performs well for answering various questions about diseases and drugs. The accuracy of the system for 500 sample questions was 83.6% [16].

Hosseini developed a question classifier for Persian open-domain QA systems. Then, they redefined some syntactic and semantic features for the Persian language and automatic extraction. The classifier results revealed that class-specific related words were the most effective feature in optimizing the classifier results among all the applied syntactic and semantic features. Eventually, the best composition of features was used to combine the headword and class-specific related words, which provided an accuracy of 85.7 and 74.4 on coarse and fine classes, respectively [8].

3. PATTERNS RECOGNITION

The present study aimed to recognize causal and explanatory relations of Persian text, which can answer “how” and “why” questions. In this section, the first phase of developing this objective (including the presence of causal and explanatory relation in sentences) was described. Hence, a pattern recognition model was developed to indicate method-effect/ cause-effect information in sentences. The adopted method used a collection of manual linguistic patterns showing the existence of a purposeful defined relationship.

In the present study, the causal and explanatory phrases were used as superclass phrases, each one referring to several relations of one class. The relations (i.e., cause, effect, and purpose) were assigned to when the causality term was used. Therefore, the explanatory term was employed for referring to relation (i.e., interpretation, explanation, evidence).

This method was based on slot filling and pattern matching of information extraction. This set used some predefined linguistic patterns with a text in natural language for extracting relation type and information of cause-effect/explanatory-effect. The patterns were created by analyzing an extracted dataset of an extensive unlabeled Persian text set called HamshahriCorpus. This corpus set is a reliable standard Persian text in 2008 and 2009 to evaluate the Cross-Language Evaluation Forum (CLEF) systems.

First, a set of patterns was made by different types of separated tokens with void space. These tokens were made easier to improve and develop a set of new patterns. A separate algorithm converted each pattern of set in sequences of letter characters and specific phrases. The method to build some patterns was presented in the following. The created patterns must be matched to cover different interpretations of sentences and syntactic structures because verbal communication is a function of linguistic apparatus achieved from semantic logic. The pattern was developed via multiple phases of reasoning approaches. The deductive and inductive steps were integrated as a cycle; these patterns were set into this cycle to create a set of 2000 general patterns.

Inductive step: As the first phase of the development process, this step comprises special observation of a sample of sentences with causal relations. In this step, special experimental patterns are prepared to determine cause and effect slots. For instance, pattern P (3) created by sentence (5) indicates that the term before is effect slot, while the following terms are shown as the cause.

P (3) R (&C) [E] AND] + C.& [

Inductive step: the output patterns of the previous phase were investigated to test them versus extracted textual segments from the corpus. Each text segment consists of causal section occurrence (which is investigated by a pattern) and a "window" of ten words after and ten words before that. A longer "window" was required in most cases because the Persian author preferred to use more significant grammatical segments and aggregation.

Three types of errors may occur after testing patterns in the inductive step. Error types were improved and investigated by running another phase of the deductive phase. These errors can be grouped as follows:

1- Undetected relation: This relation is applied when the created pattern cannot detect a causal relation in a text segment. Thus, more patterns must be added to detect the missing relation.

It is maybe better to develop a pattern to cover all intangible relations for converting that from a special to a general pattern. For instance, the previous pattern p (3) in sentence (5), which created to detect causality relation, cannot detect the provided causality relation in sentence (6) because one of the features of pattern p (3) deleted from sentence (6) is the word "because". Hence, pattern P (4) was created to retrieve the unsuccessful relation.

P (4) R (&C) [E] AND C.& [

2- Irrelevant relation: Such a relation shows a condition that a structured pattern cannot be detected a relation as a cause correctly. Domain limitation of these patterns must be limited from larger to more special with more limitations. In addition, a new pattern of void value must be added to exclude the phrases, leading to an error for correcting this error.

3- Misidentified slots: The pattern cannot complete the cause and effect slots correctly sometimes, even if a relation was correctly extracted. Resorting the patterns is an excellent solution to meet this fault and change the priority of using patterns to exploit.

For instance, the cause and their effect slots in sentence (10) cannot be filled correctly by pattern P (5). Thus, another extra pattern such as P (8) is required, which is generated and entered before pattern P (5).

P (8) R (&C) [C] (AND] (E.& [

Algorithm 1 Convert a linguistic pattern to a regular string phrase

```
1. Input: A linguistic pattern.
2. Output: The equivalent regular expression string.
3. Replace [c] and [E] symbols with “(bw+/w+b)+”;
4. Replace all pair of braces with “()?” ;
5. Replace all POS Tags (tag) with “(_+/tag_)”;
6. Replace all (/) symbols with “—”;
7. Replace all C characters with “(_+/+_)+”;
8. Replace all (Wn) symbol with “(_+/+_)1,n”;
9. If a token starts with (#) symbol
10. Add the string “( —— )?” to the beginning of the token;
11. If a token ends with (@) symbol
12. If a token starts with (&) symbol Then Retrieve the list of the words and phrases referred to by
    the token and replace it with the token as a one set of alternative strings;
13. If a token starts with $ symbol
14. Replace all w characters with “”;
15. Replace all A characters with “ ;”
16. Replace all a characters with “ ;”
17. Replace all Y characters with “ ;”
18. Replace all W characters with “ ;”
19. Replace all M characters with “ ;”
20. Replace all Q characters with “ ;”
21. Replace all y characters with “ ;”
22. Replace all N characters with “ ;”
23. Replace all C characters with
24. Replace all E characters with
25. End If
26. If a token starts with (!) symbol
27. Replace all with “ ; ”
28. Replace all with “ ; ”
29. Replace all with “ ; ”
30. End If
31. Replace all white spaces with “”;
32. Omit all previous symbols from the string;
33. Convert all Persian letters into the equivalent UTF-8 encoding characters;
34. END
```

The actions for converting designed patterns to the equivalent regular phrases are described by Algorithm (1).

This algorithm substitutes each pattern token to match with the output POS tag. Tagger generates a sequence of tagged words in tag/word form. For example, all POS tags in a pattern are set in line 3, which are replaced by a string initiating with the “b” boundary character. In addition, a “w” word character connected to the operator “+” is along with that to match with one or more events in each Persian letter. Therefore, the purposed POS tag (i.e., “tag”) is limited to the phrase “b”. The algorithm substitutes the Persian words’ symbols with the real Persian letters in lines [12-28]. In the end, all special symbols are removed by the algorithm, followed by mapping Persian characters in the pattern with the character corresponding to UTF-8 encoding.

4. PREPOSITIONS

In this section, these patterns were divided to four categories of frequently used Persian letters such as , and to meet semantic ambiguity of these phrases using the defined algorithm. First, these prepositions were tagged in the textual corpus of the present research semantically.

Several meanings of letter in the corpus: 1) by, 2) from, 3) kind and type, 4) about and related, 5) reason and cause, 6) include, 7) size and value, 8) time and period, 10) way and through, 11) into virtual or spatial range, 12) differentiate, 13) section and part, 14) piece, 15) location, 16) belonging to, 17) created and via, 18) in terms of and from, and 19) condition and how.

Several meanings of letter in the corpus: 1) time range, 2) location range, 3) about and related to, 4) virtual location range, 5) inside and into 6) into a virtual domain, 7) reason and causality, 8) present, condition, and how, 9) type, 10) when, 11) alternating, 12) size, 13) against, and 14) in comparison.

Several meanings of “” letter in the corpus: 1) expressing the relation between two or more persons or concept, 2) through, 3) by and via, 4) including and have/has, 5) with and along with, 6) condition and status, 7) against and versus, 8) kind and type, 9) time, 10) about and infield, 11) location, 12) appropriation, 13) differentiate, 12 reasons and cause, and 15) how.

Several meanings of “” letter in the corpus: 1) end of time, time interval, 2) distance, 3) value, 4) difference and types, 5) virtual distance, and 6) cause and reason.

Next, semantic patterns were defined to prepositions and Patterns consist of the syntactic category of the word (i.e., definition or meaning of the word in context). The semantic inclusion of the patterns before and after prepositions or conjunction can be extracted from the corpus and its Persian version, FarsNet. About 73 semantic patterns were defined for prepositions or conjunction because 73 different meanings were assigned semantic tags to prepositions or conjunction in the textual corpus. Also, algorithms, through their information, met the ambiguity of the letters. A sample of considered pattern to letter was described as follows:

Semantic pattern

Grammatical category: “prepositions”

Definition: “a role word which indicates a duration of a period or the beginning point of a period.”

“beginning, period”

Sample: “from today”, “From this year”, “From long ago”, “From this moment”

After semantic inclusion: “time”, “period”

Before semantic inclusion: “time”, “period”

An algorithm was developed and proposed a judgment about the words which begin with each of these letters includes an explanatory function or not. The algorithm determined the meaning of prepositions before and after words. Furthermore, the algorithm was based on the relation between semantic inclusion words as prepositions and inclusion words in entries.

Algorithm 2 It determined the type of prepositions or conjunctions in sentences.

```
A word W1 that is Preposition Or Conjunction
The Tagged sentence in which W1 appears
Stop Words List.
Output: Determination Of whether W1 Constitutes a justificatiob relation Or Caustion Relation?

If W1 Not contains in Stop words list
    Return false;
    For index = 1 To 3 For W's that appear after W1
        IF W that appear after W1 is a noun
            If Hyponymy(W) Is In Frames Then
                If Hyponymy = Or Hyponymy = Or Hyponymy = Then
                    Return True
                End If
            End If
        END IF
    Next
    For index = 1 To 3 For W's that appear before W1
        IF W is a noun
            If Hyponymy(W) Is In Frames Then
                If Hyponymy = Or Hyponymy = Or Hyponymy = Then
                    Return True
                End If
            End If
        END IF
    Next
    Return false
End If
```

In the present study, the proposed method was determined by explanatory and causal letters in Algorithm (2).

5. RHETORICAL STRUCTURE THEORY IN PERSIA QA SYSTEM

Rhetorical Structure Theory (RST) organized the text using the relation between text units as rhetorical relation. Each unit of a text has a role as core or margin, in which the cores are the significant part of a text, and margins help to cores and are minor. The rhetorical relation must be used in texts to recognize potential answers because answering to “how” and “why” questions are reasonably parts of the text related to what is being questioned rhetorically. The RST districted the parts of the text that recognize the author's primary purposes (called core) and the parts that provided complemental items called margin. Therefore, this is an appropriate tool to analyze text deductive paragraphs. The computation performance was increased by incorporating the relations that considered a specific sentence at a time. Furthermore, each assumed RST relation with the heuristic score affected the text parser and created an appropriate tree, and avoided hybrid produce the trees and any computational explosion.

Two models were considered in the current work for developing the suggested method. The first one was a pattern recognizer that divided the text into Elementary Discourse Units (EDUs) sentences

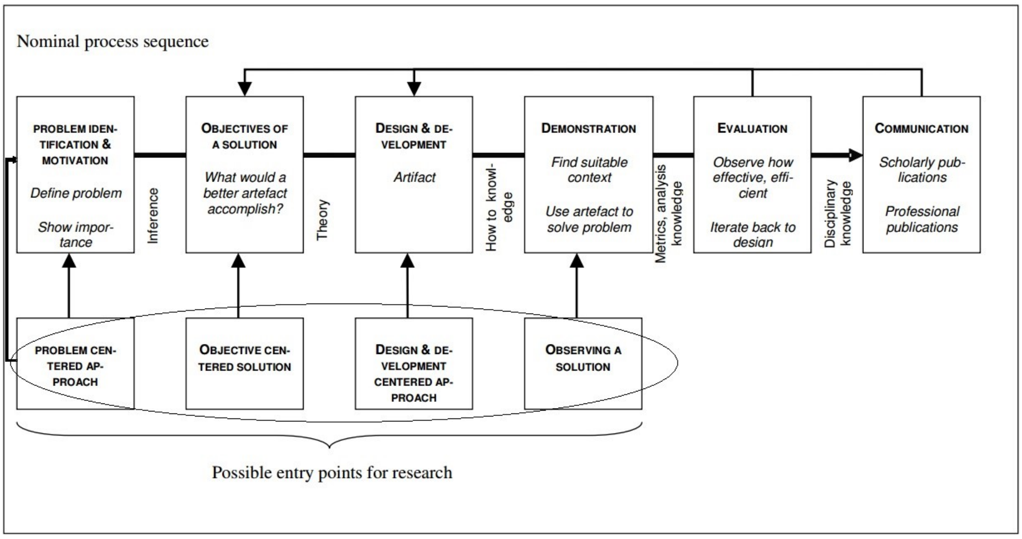

Nominal process sequence
PROBLEM IDENTIFICATION & MOTIVATION
Define problem
Show importance
Inference
OBJECTIVES OF A SOLUTION
What would a better artefact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to knowledge
DEMONSTRATION
Find suitable context
Use artefact to solve problem
Metrics, analysis knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM CENTERED APPROACH
OBJECTIVE CENTERED SOLUTION
DESIGN & DEVELOPMENT CENTERED APPROACH
OBSERVING A SOLUTION
Possible entry points for research

Figure 1: Flowchart of the model proposed in this research

and provided semantic relations using linguistic patterns. The second model was a Text Parser on related sentences with relation slots defined by the first model and setting rhetorical relation between adjacent text units with at least one sentence.

A primary sample was implemented to evaluate the proposed concept of the Persian QA system, which applied the presented framework tax law and documentation. In the present study, the “how” and “why” questions were described as a question sentence

6. METHODOLOGY

The methodology followed in the present study is based on the design science research (DSR) pattern. Hevner claimed that this approach is proper for IS researches, and the DSR explores solutions to “important and related problems to business” [6] [7]. Also, Peffers provided the DSR model, covering the entire research project from motivation to relation [12]. This process of the model is presented in Fig. 1. that the noun “for what” and “how” occurred in the primary position. The overall architecture of the design QA system presents in Fig. 2. The secondary sources (such as the Tax law database, Books, Rules, Thesis, and the Internet) were used in the present study to extract tax questions and texts. The information retrieval and structure detection algorithm of texts and sentences was used to analyze data and design the system. The sample software was designed using Python programming language and the SQL Server 2016 database.

7. SYSTEM DESIGN AND IMPLEMENTATION

Fig. 1 briefly describes the overall class diagram of the research system. The major class was “QuestionAnswering”, which distributed functions in 3 packages. In the first package, the main class was “PatternRecognizer”, which used the created set of linguistic patterns to find the inter-sentence relation in the text. In this package, “Tokenizer” and “POS tagger” were defined for recognizing the defined patterns. In the other package, the primary class was “TextParse”, which analyzed the previous package’s tagged sentences and used “DiscourseMarkers” for detecting between sentence relations. In the third package, the “answerFinder” class prioritized the “gettingkeywords” (which recalled the “Stemmer”) and “Tokenizer” classes (which enabled the “Similarity” class) to find more appropriate answers.

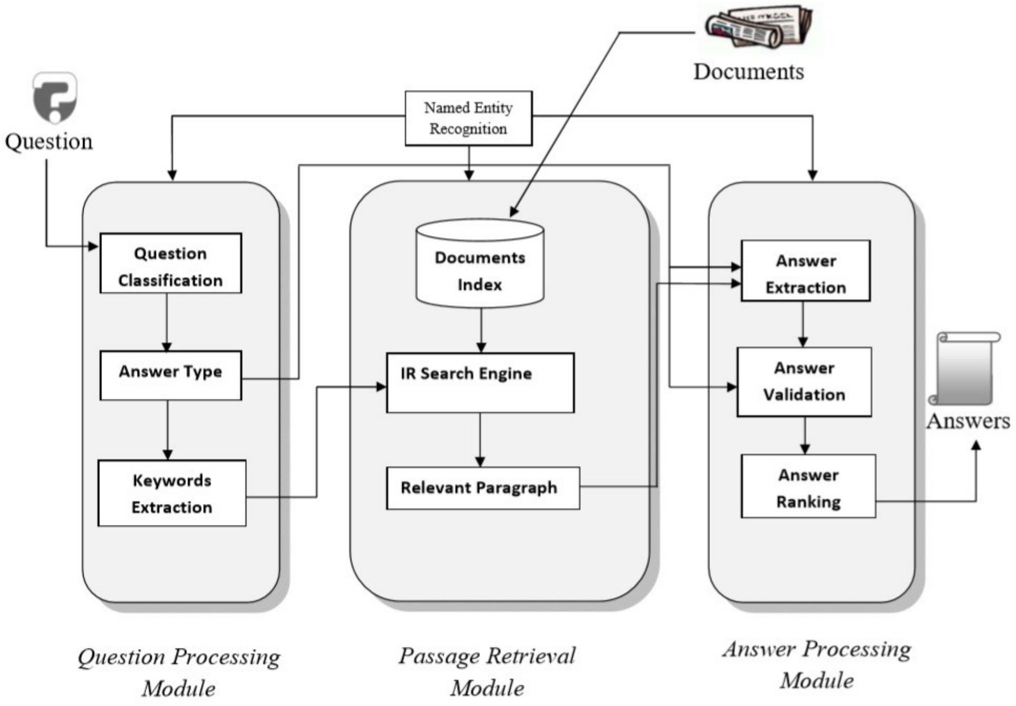


Figure 2: System design architecture

8. System implementation

The system interface was designed using the C# programming language in the .Net framework. The samples of produced relation in asking a question related to the text present in Figs. 4, 5, and 6. The evaluations were classified into two steps. The first one describes how the constructed linguistic patterns can recognize the inter-sentence relation and the path of these relations. In the second phase, the overall QA system performance was evaluated. Based on the evaluation of the second phase, a strategy similar to Verberne was used for evaluating the fitness of selected text units by system as a candidate answer about "how" and "why" questions [17].

All performed experiment was based on a set of researches from Persian Hamshahri Corpus. This set covered 160,000 documents in the Persian language and a wide range of texts. This corpus covered 50 different topics and texts written by professional authors. Hamshahri corpus is provided by the Database Research Group of Tehran University supported by the Iran Telecommunication Research Center. The researches were collected from two health and science/technology categories containing 2138-485 words individually. In the experiments, five independent candidates with native Persian language participated. All participants are educated, three were linguistics doctorate, and two were communication experts. In both experiments, the evaluation was performed by comparing the generated output by the system versus the judgment of these participants.

9. Linguistic pattern evaluation

In this section, the experiment was performed in two phases. In the first phase, only the linguistic patterns were used to recognize inter-sentence relations. In the second phase, the prepositions were mentioned.

Also, the 11 texts were segmented according to the complete break of a sentence manually, i.e., end of the sentence with punctuation marks, and the results were 415 sentences. Three participants read

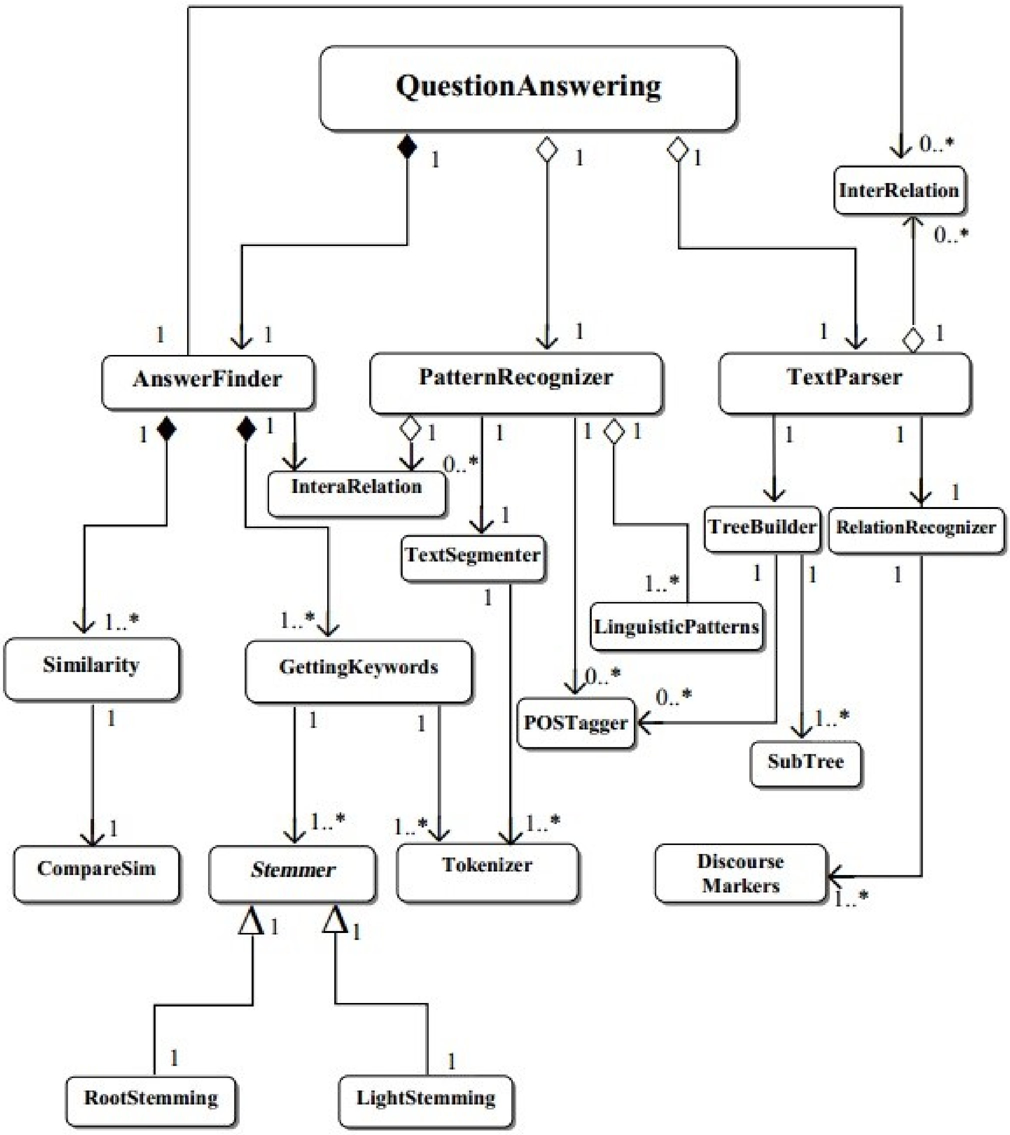


Figure 3: Overall diagram class of QA system

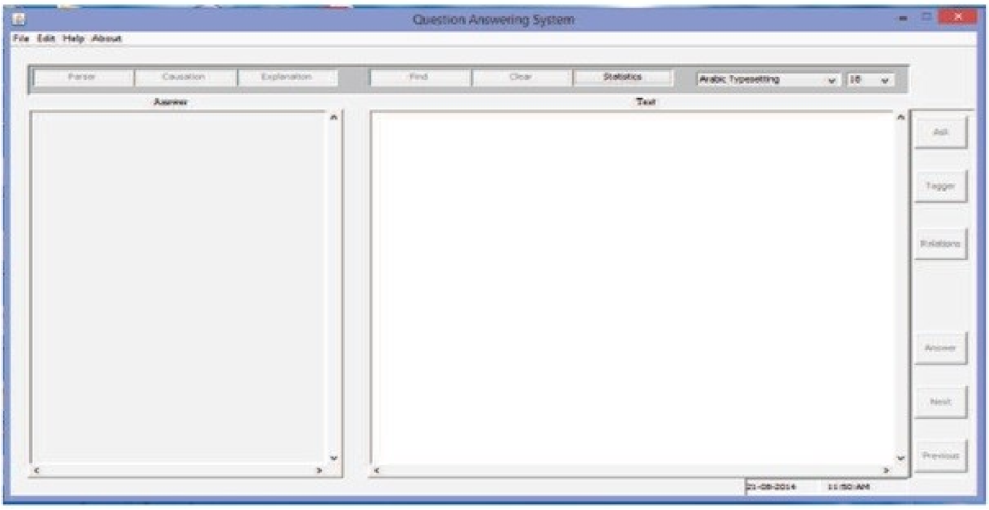
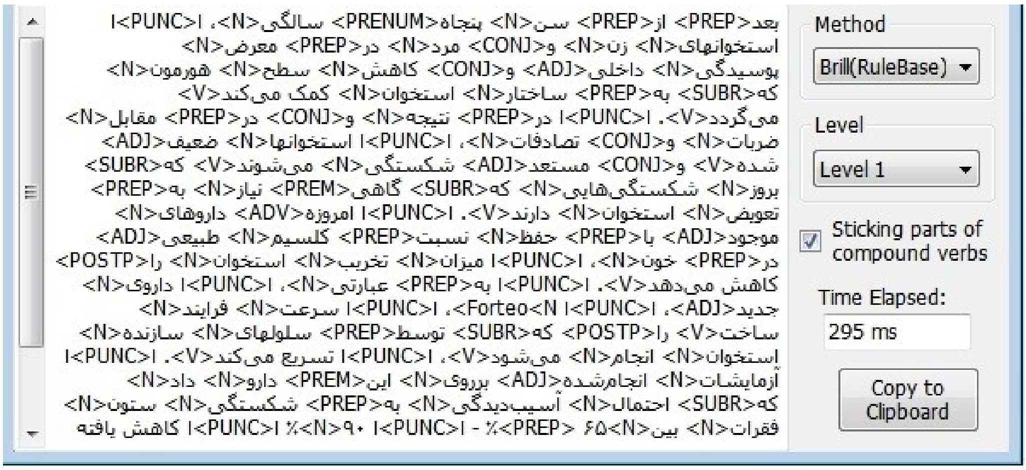


Figure 4: The major interface of the QA system



بعد<PREP> از<PREP> سن<N> پنجاه<PRENUM> سالگی<N>، ا<PUNC>ا
استخوانهای<N> زن<N> و<CONJ> مرد<N> در<PREP> معرض<N>
پوسیدگی<N> داخلی<ADJ> و<CONJ> کاهش<N> سطح<N> هورمون<N>
که<SUBR> به<PREP> ساختار<N> استخوان<N> کمک می‌کند<V>
می‌گردد<V>. ا<PUNC>ا در<PREP> نتیجه<N> و<CONJ> در<PREP> مقابل<N>
ضربات<N> و<CONJ> تصادفات<N>، ا<PUNC>ا استخوانها<N> ضعیف<ADJ>
شده<V> و<CONJ> مستعد<ADJ> شکستگی<N> می‌شوند<V> که<SUBR>
بروز<N> شکستگی‌هایی<N> که<SUBR> گاهی<PREM> نیاز<N> به<PREP>
تعویض<N> استخوان<N> دارند<V>. ا<PUNC>ا امروزه<ADV> داروهای<N>
موجود<ADJ> با<PREP> حفظ<N> نسبت<PREP> کلسیم<N> طبیعی<ADJ>
در<PREP> خون<N>، ا<PUNC>ا میزان<N> تخریب<N> استخوان<N> را<POSTP>
کاهش می‌دهد<V>. ا<PUNC>ا به<PREP> عبارتی<N>، ا<PUNC>ا داروی<N>
جدید<ADJ>، ا<PUNC>ا Forteo<N>، ا<PUNC>ا سرعت<N> فرایند<N>
ساخت<V> را<POSTP> که<SUBR> توسط<PREP> سلولهای<N> سازنده<N>
استخوان<N> انجام<N> می‌شود<V>، ا<PUNC>ا تسریع می‌کند<V>. ا<PUNC>ا
آزمایشات<N> انجام‌شده<ADJ> برروی<N> این<PREM> دارو<N> داد<N>
که<SUBR> احتمال<N> آسیب‌دیدگی<N> به<PREP> شکستگی<N> ستون<N>
فقرات<N> بین<N> ۶۵<PREP>٪ - ا<PUNC>ا ۹۰<N>٪ ا<PUNC>ا کاهش یافته

Method: Brill(RuleBase)

Level: Level 1

☑ Sticking parts of compound verbs

Time Elapsed: 295 ms

Figure 5: The screenshot of Pos tag of entered text

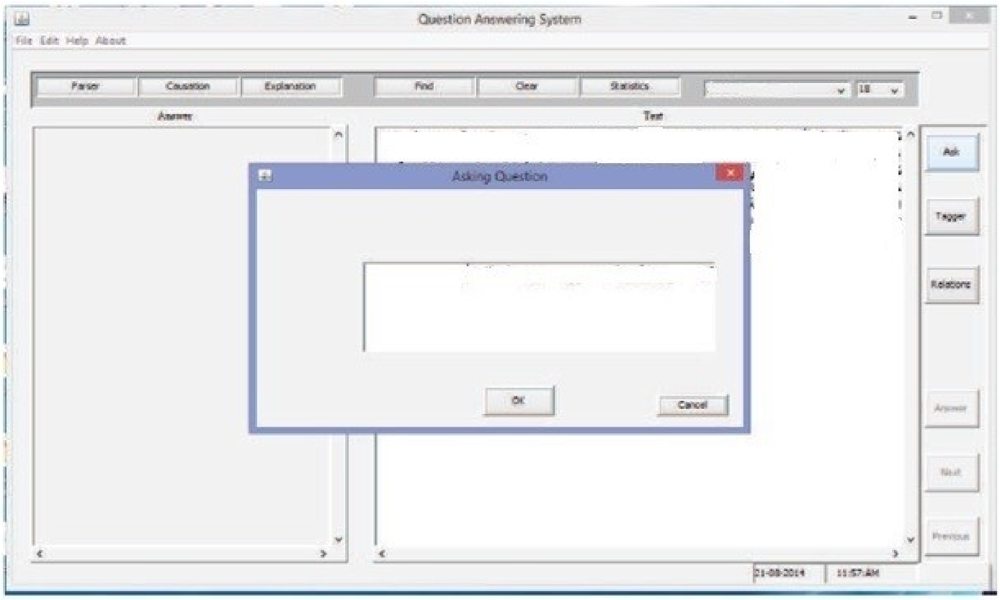


Figure 6: Question form provided by the user

the sentences and showed the explicitly causal relation in each sentence with cause and effect slots. As a result, 240 causality relations were obtained. Then, the pattern recognizer was performed to extract similar information.

The performance criteria were recall and precision. The recall value was the ratio of recognized relation by participation and pattern recognizer. In other words, the recall was equal to the recognized value by system division on the recognized value by the participation manually. The precision was the ratio of recognized relation by pattern recognizer, which also recognized participations. In other words, the precision value was equal to the recognized value by participation manually division on the recognized value by system.

Table 1: The recognized relations to a health text without prepositions

	Manual	System	Recall
Text 1	19	12	0.63
Text 2	32	14	0.44
Text 3	33	10	0.3
Text 4	18	11	0.61
Text 5	11	4	0.36
Total	113	51	0.45

Table 2: The recognized relations in a health text with prepositions

	Manual	System	Recall
Text 1	19	16	0.84
Text 2	32	28	0.87
Text 3	33	23	0.70
Text 4	18	13	0.72
Text 5	11	8	0.73
Total	113	88	0.78

Tables 1 and 2 represent the number of recognized relations by participating in each health text with and without adding a preposition algorithm. In the second column, the number of relations recognized by the pattern recognizer correctly is shown.

Table 3: The recognized relations to scientific text without the preposition algorithm

	Manual	System	Recall
Text 1	18	5	0.27
Text 2	18	12	0.66
Text 3	27	15	0.55
Text 4	7	5	0.71
Text 5	24	13	0.54
Text 6	33	24	0.73
Total	127	74	0.58

Table 4: The recognized relations to scientific text with the preposition algorithm

	Manual	System	Recall
Text 1	18	16	0.89
Text 2	18	15	0.83
Text 3	27	24	0.88
Text 4	7	6	0.86
Text 5	24	17	0.71
Text 6	33	29	0.88
Total	127	107	0.84

Tables 3 and 4 indicate the same information for science and technological text. Pattern recognition achieved the highest overall recall values of 78% and 84% for the health text and science and technological text, respectively.

Table 5: Precision, Recall, and F-measure relation to a health text without prepositions

	Recall	Precision	F-score
Text 1	0.63	0.95	0.76
Text 2	0.44	0.97	0.61
Text 3	0.30	0.91	0.45
Text 4	0.61	0.98	0.75
Text 5	0.36	0.94	0.52
Total	0.45	0.95	0.61

Table 6: Precision, Recall, and F-measure relation to health text with prepositions

	Recall	Precision	F-score
Text 1	0.84	0.86	0.85
Text 2	0.87	0.84	0.85
Text 3	0.70	0.74	0.72
Text 4	0.72	0.88	0.79
Text 5	0.73	0.67	0.70
Total	0.78	0.8	0.79

Tables 5 and 6 present the Precision, Recall, and F-measure relation to the health text and prepositions algorithm.

Table 7: Precision, Recall, and F-measure relation to the scientific text without prepositions

	Recall	Precision	F-score
Text 1	0.27	0.93	0.42
Text 2	0.66	0.96	0.78
Text 3	0.55	0.93	0.69
Text 4	0.71	0.95	0.81
Text 5	0.54	0.94	0.69
Text 6	0.73	0.85	0.76
Total	0.58	0.93	0.71

Table 8: Precision, Recall, and F-measure relation to the scientific text with prepositions

	Recall	Precision	F-score
Text 1	0.89	0.80	0.84
Text 2	0.83	0.75	0.79
Text 3	0.88	0.75	0.81
Text 4	0.86	0.88	0.87
Text 5	0.71	0.74	0.72
Text 6	0.88	0.71	0.79
Total	0.84	0.76	0.80

Tables 7 and 8 show similar actions to health text and prepositions algorithm. The F-score was calculated by equation (1). The F-score is always a number between the values of recall and precision.

$$F = \frac{2 * precision * Recall}{Precision + Recall} \tag{9.1}$$

Investigating the relation recognized by the pattern recognizer in the second phase of experiments revealed that 30 relations of the set (67%) were created due to the particular type of words not contained in the patterns list. Some of these words were used to indicate this relation rarely. The recognizer cannot recognize another set of relations (15 relations) due to the unexpected structure of the sentence, which covered 33% of lost relations. This type of relation was indicated implicitly, and the required overall knowledge was deduced to recognize such relations.

10. Results and Discussion

A normal pipeline question-answering (QA) system includes question-, document-, and answer-processing phases. Most studies on the Persian language have classified questions and answers and search engine design. Questions can be categorized using rule-based (manual), machine learning, and hybrid approaches.

It is noteworthy that most research studies on question and answering systems only address solutions to question and answering classification while not considering the structure of the Persian language and linguistics.

We believe that the most difficult and challenging type of questions in Persian question and answer systems are questions that seek the cause and why of an issue. As a result, using the same methods to answer different types of questions with one solution simply categorizing the questions and retrieving the answer or document cannot be complete.

Answering the questions of why and how is possible with the conventional methods of classifying and retrieving the answer, but the variety of structural relationships of languages, especially in Persian, will have a great impact on the meaning and concept of the returned answer.

we investigated different studies which tackled mining causation in texts written in languages other than Persian. A number of these studies used hand-coded pattern and specific knowledge bases. Other systems employed machine learning approaches in order to automatically construct syntactic patterns.

Researchers employing machine learning techniques made use of knowledge resources available for the language they addressed, e.g. (large annotated corpora, WordNet, Wikipedia etc.). Such resources provide externally verified analyses of POS and constituency, and are invaluable for those desiring to evaluate and train models that involve statistical component. Given a similar corpus of Persian texts annotated with Causal and Explanatory relations, it should be possible to automatically acquire patterns.

In fact, it is challenging to capture the syntactical arrangement of many of the causative connectors. Accordingly, machine learning approaches followed in research presented in Section 1, could not be applied in this study due to lack of large quantities of annotated data. Handcrafting in order to construct linguistic patterns able to indicate semantic relations within sentences is therefore still necessary.

As the Pattern Recognizer model obtained a maximum overall recall of 81% we conclude that using the linguistic patterns boosted with the justification particles algorithms will be effective for identifying intrasentential information . Furthermore, the extracted linguistic patterns reflect strong relation indicators and constitute a useful feature in the future for systems adopting machine learning techniques in acquiring patterns that signal causation and explanation.

In this research, the aim is to produce a system in accordance with software standards. The methodology in this research(Figure 1) has been selected in such a way that the usual model of question and answer systems (Figure 2) has been considered. The output is considered as an artifact that is grown in accordance with analysis and validation and placed in a modification cycle to produce the final product.

This is one of the main differences between this article and others. The Q&A model is adapted to the concepts related to software standards and the class of diagrams is fully described.

As the Pattern Recognizer model obtained a maximum overall recall of 81% we conclude that using the linguistic patterns boosted with the justification particles algorithms will be effective for identifying intrasentential information . Furthermore, the extracted linguistic patterns reflect strong relation indicators and constitute a useful feature in the future for systems adopting machine learning techniques in acquiring patterns that signal causation and explanation.

Although in some articles the accuracy of the proposed methods is higher than the accuracy of the present article, but it should be borne in mind that in this article our emphasis has been on the accuracy and precision of answering the questions why and how by the system. The one-word and short-answer questions discussed in the output system are not included for testing and validation. As explained earlier, questions can be classified using the three mentioned approaches. Manual question categorization seeks to conform questions to hand-made rules to determine the type of answer to the question. It is of note that writing down such rules is monotonous and the final system is frequently particular. Besides, this type of system is scarce since the overall performance does not even resemble machine learning-based approaches, which are entirely diverse and some of which attempt to demonstrate questions as a tree. However, our proposed method relies on rhetorical structure theory (RST), which integrates natural language discourse. RST structures are documented hierarchically by breaking the content into (sub-) clauses called Elementary Discourse Units (EDUs). Then, EDUs are clinked to build a binary discourse tree. RST differentiates between a nucleus and satellite, which convey primary and ancillary information, respectively.

The some mentioned studies[8] emphasized answering questions to which the answer is a fact. The articles do not focus on the specific semantic structure of the Persian language and the effect of prepositions on the meaning of the questions, but rather focus on clearing and ultimately categorizing and matching questions and answers.

In other article[16] that have been done in Persian language and with the approach of questions with non-factoid QA, the classification of the questions is based on the classification algorithm and without considering the type of semantic connections of the answers. The accuracy of this study is 82.29, but it considers all types of non-factoid questions and does not focus on the semantic relationship between questions and answers.

In comparison, our study focused on answering why and how while highlighting the semantic aspect of the answer to the question. The Persian articles were distributed between five persons for evaluating the performance of the QA system. They were asked to read some texts and regulate “how” and “why” questions to find the answer in the text and to provide related answers for each question. Finally, 90 QA pairs were achieved, including 70 questions for why and 20 questions for how cases.

As discussed earlier, non-factoid QA has received less attention in QA systems than fact-based QA due to the required linguistic knowledge for addressing such questions. Nonetheless, many researchers have become interested in adopting new techniques to handle explanation and reasoning questions in recent years. The proposed system was implemented on 90 collected questions. Then, the achieved answers by systems were compared to the users’ formulated answers. The correct answer was considered if the answer matched the user’s answer. This system can obtain the correct answer for 61 questions, which results in a 68% recall rate. Table 9 provides an overall review of system results.

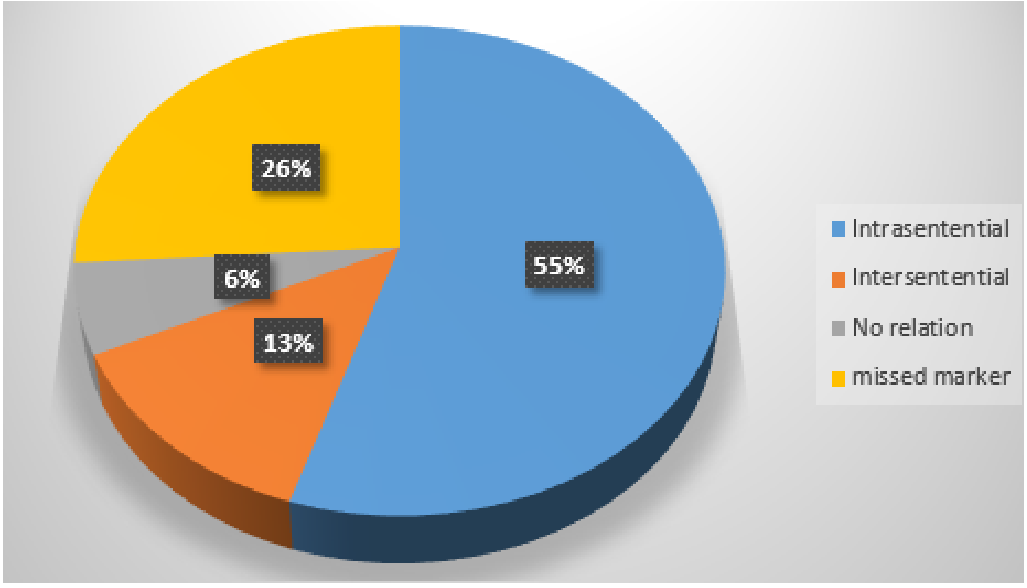


Figure 7: The distribution of the questions test

Table 9: QA system results

	number of questions	Percentage of Total questions
number of questions	90	100
Correct answer count	61	68
Wrong answer count	29	32

Note. QA is the question-answering system.

The system could not find a correct answer to two groups of questions. The first group consisted of the questions without a direct relation between the text unit of the question and the answer (5 questions, 18% of which were unanswered) and were connected to the answered units with an implicit relation in the text. The second group included 24 questions, 82% of missing questions, in which Text Parser or Pattern Recognizer does not support the linguistic parts. Questions in this category are connected to the answers spans with relations expressed implicitly in text.

The test data were collected through elicitation which implies that questions might have been influenced by the same linguistic cues used by the text producers. This results in lexical overlap more than one would expect for natural questions. It is important to remember, however, that the ultimate goal of question answering systems is to find answers in vast amounts of information which users might not have access to. Future work should be dealing with questions formulated independently of a specific text. To reach this goal consideration must be given to the query expansion techniques. Figure 7 illustrates the distribution of the questions answered correctly (green colored partitions) together with the failed questions (red colored partitions). Nearly 55% of the questions were answered correctly based on the indication of intrasentential relations, whereas correct answers for 13% of the questions correlate to the presence of rhetorical relations between sentences. In this study, we built our discourse parser on top of the output obtained from the Pattern Recognizer which is sentences that are already annotated with intrasentential relationships. To fulfil this goal , we developed the Text Parser model that would approach text from a discourse perspective. The Text Parser is meant to break away from the sentence limit imposed on the Pattern Recognizer and emphasize the strategies employed above these limits to hold the whole text together as a unit. Furthermore, the Text Parser is led by a set of heuristic scores to avoid any computational explosion.

11. CONCLUSION

The primary motivation of the present study was to design simple techniques to find the answers to understand and work the “why” and “how” questions easily and quickly. It is predicted that a slot of Persian QA systems was filled. Hence, two analytical models were provided as Pattern Recognizer and Text Parser, which were created with high precision and low complexity. The present study aimed to recognize semantic relations in Persian sentences. Based on morphological and syntactic features, a collection of linguistic patterns was generated. The pattern recognizer model was used for applying the patterns to extract method-effect and cause-effect data. These data are critical to the QA system to answer the “how” and “why” questions. Furthermore, an algorithm was proposed to increase the pattern recognizer’s effectiveness with determining the explanatory/causal role of rationalization letters, as another concern of this study.

The morphological scheme of the Persian language is relatively complicated due to morphological changes and agglutinating phenomena. In addition, the Persian language is very glamorous such that a large part of its words is originated from the roots bounded by many suffixes or prefixes, or both of them. These suffixes and prefixes can be related to any Persian word groups, like verb, noun, or adjective, doubling the Persian language’s work in QA systems.

In the feature studies, the researchers can use machine learning techniques on existed knowledge resources. In addition, they can obtain automatic patterns by a similar corpus of marginalized Persian texts with causal and explanatory relations.

References

[1] E. Blancol, N. Castell and D. Moldovan, *Causal Relation Extraction*, In Proceedings of the International of Conference on Language Resources and Evaluation, LREC, Morocco. (2008) 310-313.

[2] Y. Boreshban and S.A. Mirroshandel, *A novel question answering system for religious domain in Persian*, Journal of ELECTRONIC INDUSTRIES. 8 (2017) 73–88.

[3] E. Breck, J. Burger, L. Ferro, D. Hous, M. Light and I. Mani, *Another Sys Called Qanda*, In Proceddings of the Ninth Text REtrieval conference, NIST Special Publication 500-246, Maryland,. (2000) 369-379.

[4] K. Choi, R.M. Pacana, A.L. Tan, J. Yiu and N.R. Lim, *A Question Answering System that Performs Evaluations and Comparisons on Structured Data for Business Intelligence in Biotechnology*, Journal of Uncertainty Reasoning and Knowledge Engineering. 1 (2011) 137–140.

[5] J. Herrera, D. Parra and B. Poblete, *Social QA in non-CQA platforms*, Journal of Future Generation Computer Systems. 105 (2020) 631–649.

[6] A. Hevner, S. March, J. Park and S. Ram, *Design Science in Information Systems Research*, MIS Quarterly 28, (2004) 75–105.

[7] A. Hevner and S. Chatterjee, *Design Science Research in Information Systems*, Springer Science and Business Media. (2010) 195-208.

[8] H. Hosseini, *Question Processing for Open Domain Persian Question Answering Systems*, M.Sc. Thesis ,Sharif University of Technology, Department of Languages and Linguistics. (2016).

[9] C.S.G. Khoo, S. Chan and Y. Niu, *Extraction Causal Knowledge from a Medical Database Using Graphical Patterns*, In Proceedings of 38th Annual Meeting of the ACL, HongKong. (2000) 336-343.

[10] A. Mollaei, S. Rahati-Quchani and A. Estaji. *Question classification in Persian language based on conditional random fields*, International conference on computer and knowledge engineering. (2012) 295–300.

[11] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lascatusu, A. Novischi, A. Badulescu and O. Bolohan, *LCC Tools for Question Answering*, In Proceedings of the Eleventh Text REtrieval Conference, NIST Special Publication 500- 251, Maryland. (2002) 386–395.

[12] K. Peffers, T. Tuunanen, A. Marcus, Rothenberger and S. Chatterjee, *A Design Science Research Methodology for Information Systems Research*, Journal of Management Information Systems. 24 (2008) 45–77.

[13] M. Razzaghnoori, H. Sajedi and I. Khani Jazani, *Question classification in Persian using word vectors and frequencies*, Journal of Cognitive Systems Research. 47 (2018) 16–27.

[14] M. Sarrouti, S. Ouatik and E.L. Alaoui, *SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions*, Artificial Intelligence In Medicine. 102 (2020) 631–649.

[15] E. Sherkat and M. Farhoodi, *A hybrid approach for question classification in Persian automatic question answering systems*, International eConference on computer and knowledge engineering. (2014) 279–284.

[16] H. Veisi and H. Fakour Shandi, *A Persian Medical Question Answering System*, International Journal on Artificial Intelligence Tools. 29 (2020) 2050019.

[17] S. Verberne, *Paragraph Retrieval for Why-question Answering*, In Proceedings of the 30th Annual International ACM SIGR Conference on Research and Development in Information Retrieval, New York. (2007) 922–927.

[18] Y.F. Wang and S. Petrina, *Using Learning Analytics to Understand the Design of an Intelligent Language Tutor-Chatbot Lucy*, Journal of Advanced Computer Science and Applications. 4 (2013), 124–131.

[19] Z. Yang, Y. Li, J. Cai and E. Nyberg, *QUADS: Question Answering for Decision Support*, In proceedings of SIGR 2014: the Thirty-seventh Annual Internations ACM SIGIR Conference on Research and Development in Information Retrieval, USA. (2014) 375–384.

[20] N. Zulkarnaina and F. Mezianea, *Ultrasound reports standardisation using rhetorical structure theory and domain ontology*, Journal of Biomedical Informatics. 1 (2019) 100003.