Natural Language Understanding for Grading Essay Questions in Persian Language

Iman Mokhtari-Fard
Department of Computer Engineering
University of Jahad
Shahrekord, Iran
iman@jdchb.ac.ir

***Abstract*—** many intelligent systems are intended to communicate with users through natural language. Understanding the natural language by the computer is one of the most essential operations in natural language processing. One of the applications of natural languages is in the exams having essay questions. The objective of this paper is to propose a method for designing an examiner machine and creating an intelligent evaluator to grade the users' given answers to the essay questions. Algorithms such as "phrase structure" are weak at natural language processing in "free word order languages" such as Persian. The recommended method in this paper is based on "dependency grammar" and is applicable for various natural languages. Reduction in evaluation time and increase in the accuracy are advantages of the proposed method.

Keywords- Artificial Intelligent , Natural language understanding , Essay Questions Examiner

1 Introduction

The ability to comprehend the natural language is one of the most valuable features of intelligent systems. Linguistic studies and designing computer algorithms are required for achieving this feature. Many studies have been conducted in this area but only the English language has been studied in most of the cases. English language is more feasible for designing algorithm compared to other languages such as German, Chinese, Persian, Arabic, etc due to its regulated structure and lack of free word order characteristic. Thus, it is essentially significant to have an algorithm which can be used for other languages as well as English.

The methods of representing the sentence grammar have played an important role in the advent of natural language processing algorithms. "Phrase structure" and "dependency grammar" are two principal methods for syntactic representation. Dependency grammar algorithm is applicable for free word order languages unlike the phrase structure method. Thus, we have applied dependency grammar and characteristics of Persian language for implementation of the algorithm.

In computer science, different methods can be applied for displaying the data. These methods include graph, Semantic networks, and use of objects. In this paper, we will represent the natural language in the form of objects with the aid of dependency grammar. A system capable of effectively understanding the natural language can be applied in different usages of natural language including summarization, text categorization, and text translation and so on. Following depiction of understanding procedure of natural language by the objects, this paper proposes a method for application in the intelligent evaluation system; the results of the recommended method are investigated in Persian language.

Syntactic representation methods are included in section II; variety of applied questions in the evaluations will be presented in section III. The proposed method is investigated in section IV. Section V deals with introducing the stages of converting texts into objects. Subsequently, answer acquisition and its evaluation procedures are analyzed in

chapter VI. In chapter VII, besides introducing the Persian language, the results of the recommended method will be presented for the sample text in Persian language, and finally section VIII will incorporate the conclusions and future activities.

2 Syntactic representation methods

Syntactic representation methods are crucially influential in natural language processing. The texts are analyzed during the stages of syntactic representation of natural language and the meaning of the text is obtained subsequently [1]. Therefore, it is particularly significant to be familiar with different syntactic analysis method. Among the syntactic representation methods, dependency grammar and phrase structure algorithms are more remarkably important [2].

1.1 phrase structure

This approach is vitally important as one of the primary methods proposed by Chomsky [3] based on which the Context-free Grammar (CFG) is operated. Every sentence in context-free grammar system consists of several phrases and each phrase is composed of a set of words. These languages are also referred as "formal languages". The grammar written based upon the formal language is called "Generative Grammar" [4].
Context-free grammars comprise a set of rules and symbols; the symbols, in turn, are divided into terminal and non-terminal symbols [2].

1.2 Grammar Representation based on the Dependency

This method was innovated by Teniere [5] as a novel approach in modern linguistics. In this method, the syntactic structure comprises words which are related to each other through asymmetrical dual relationships [6]; these relationships are called "dependency". In this representation method, it is emphasized that every sentence has a central verb, and the sentence structure can be determined using the central verb and the type and number of its mandatory and optional complements. In other words, this method rejects the former procedures which used to emphasize on division of the sentences into subject and predicate [7]. Each verb imposes specific states of dependencies; the syntactic capacity is one of the most significant concepts in the dependency grammar. The fundamental structure of a sentence is determined based on its central verb [8].

1.3 Comparison of methods

One of the major differences between the dependency and generative methods arises from the value or position they consider for the subject of the sentence. The generative grammar, from the beginning, divides the sentence into main parts of noun phrase (subject) and verb phrase (predicate); it actually follows the Aristotle logic of sentence analysis regarding the subject as one of the principal parts of the sentence just like the verb. Engel believes [9] the sentence division into subject and predicate mainly exhibits the information structure and the distribution of the new or old data rather than representing the syntactic structure of the sentence.

On the other hand, the phrase structure grammar is not suitable for languages having free constituent order. The dependency grammar is an appropriate candidate for this purpose due to its structure [2].

2 Evaluation system of users' answers

The evaluation systems are able to ask the questions from users in different ways. The different approaches for posing the question comprise following methods [10]:

- Multiple-Choice Items

Multiple-choice items include a set of responses. For each question, you must choose the best response.

- True/False Items

With true/false items, you are given a sentence to read, and you must decide whether the sentence is true or false.

- Fill-in-the-Blank

Fill-in-the-blank items are sentences with key words missing. You must fill in the missing word or phrase. You may be given a list of answer words to choose from.

- Short-Answer Questions

Short-answer questions are items that are answered by writing a few words or sentences.

- Essay Questions

An essay is writing that you do in response to a question or prompt. Essay questions test whether you have a deep understanding of your subject.

For Evaluating short-answer or essay questions, the natural language processing is required due to necessity of analysis. Many examinations including TOEFL and PhD entrance exams in many countries include essay questions. Furthermore, contests with essay questions can be held via SMS in which it may take a lot of time to evaluate all the answers because of the large number of participants; human errors might also occur in these evaluations. It is very effective to enjoy a system having the ability to investigate the answers intelligently and calculate the grade acquired for every person.

3 Recommended Method

In the proposed algorithm, the questions and correct answers are separately received by the system in the form of natural language. Then, the accurate answer for each question is converted into objects which represent the object-based representations of the input texts. In the subsequent step, the user inputs the relevant answer for each question through GUI[1]. The system converts the input text for each question into distinctive objects. For analyzing the accuracy level of the answers, the created objects are compared with each other and the answering grades are calculated.

The procedures of designing the stages were carried out in a way that can be used for different natural languages; in other words, not applicable only for a specific language. However, in the conducted researches, the procedures were studied only for Persian language which assumes a free word order.

All stages of the system have been illustrated in figure 1; the procedures start from data input in the form of natural language and continue up to analyzing the answers and calculating the users' grades:

[1] Graphic User Interface

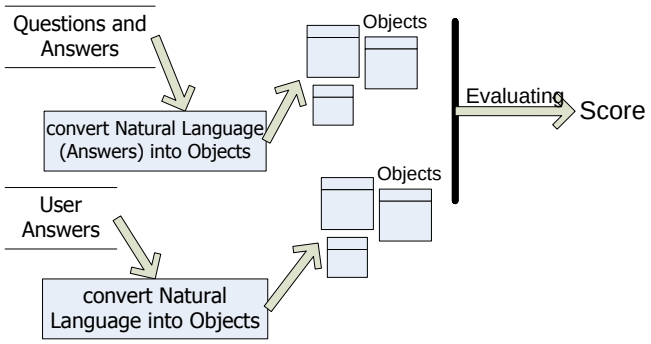


Fig. 1. System stages from receiving the questions and correct answers in natural language up to the evaluation of users' answers and calculation of grades.

According to above figure, the recommended system consists of two main phases; the questions and correct answers are received in the first phase; then, the users' answers are acquired, evaluated and graded in the second phase.

4 Procedures of converting texts into objects

We have used objects for understanding the natural language in the proposed method. Accordingly, each time a distinctive layer yields the object-based structure by acquisition of natural language. In the object-based representation, several objects might be obtained from the input text, and some relations are established between the objects whenever necessary. For converting the input text into objects, the stages are conducted as illustrated in figure 2.

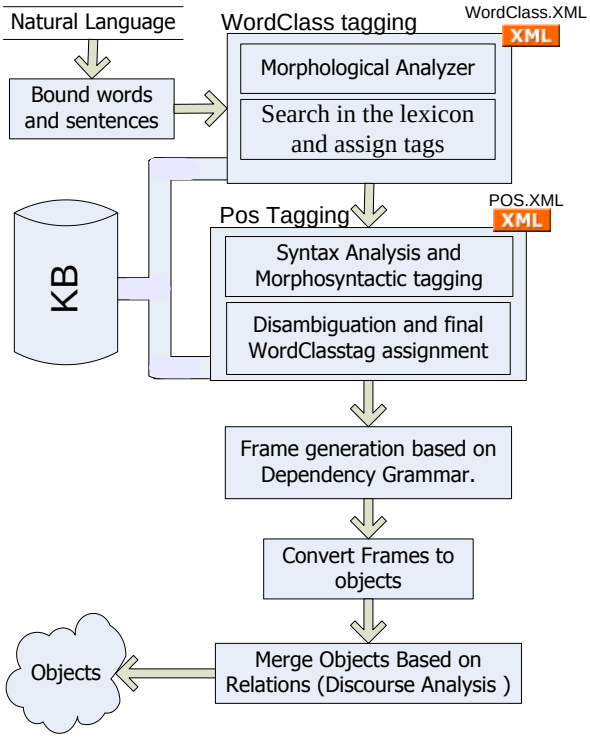


Fig. 2. Procedures of converting texts into objects

The stages illustrated in the figure are performed as follows:

4.1 Generation of initial objects

4.1.1 Word Class and Pos Tagging:

Initially, the whole word (with affixes[2]) is searched in the database. It must be noted that only the word stem has been stored so as to reduce the size of the database. In the second part, based on the linguistic rules, the word is investigated in terms of having affixes; it is then analyzed and assigned the appropriate tag if necessary.
In this stage, every word or phrase is assigned a tag based on the linguistic characteristics. If a word has been assigned several tags in word class tagging stage – for example the word "مردم" in Persian language, considering the multiple-meaning property, can have a tag as the word "مُردَم" (I died) and also another tag as the word "مَردُم" (people). Then, as this word is investigated simultaneously and in association with other words in this stage, the multiple-tag words can be disambiguated according to the word function in the sentence and the adjoining words in the text. The result of this stage is saved in Pos.xml file.

4.1.2 Frame Generation

A frame is generated by separately tagging for each sentence. Each frame includes slots to be filled by sentence constituents. The frame slots vary for every sentence and depend on the verb of the sentence. Dependency grammar characteristics are applied in this part. For instance, the verb "eat" will need an "object" dependent and a "subject" dependent (What was eaten?) (Who ate it?). Thus, the frame of the verb "eat" has got two slots. For example, for the sentences "My name is Sanova. I work in the university", two frames will be generated as indicated in figure 3:

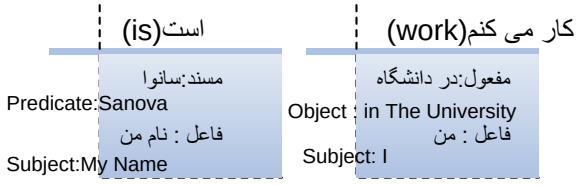


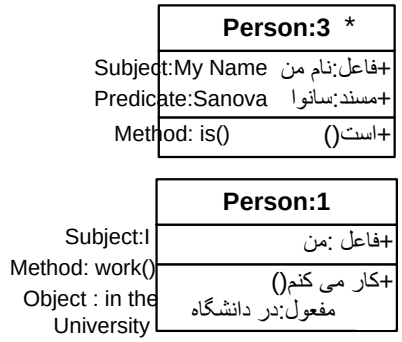
Fig. 3. The frames obtained from sample sentences

4.1.3 Converting frames into initial objects

In Persian language, every sentence has a specific person based on its verb. This person can be "first person/ second person/ third person" singular or plural. In other words, there are totally 6 persons. Accordingly, an initial object is created for every frame. The object name is a number from1 to 6 (1: for first person singular, 2: for second person singular, etc).
There exist some properties and a method inside the initial object. The method in the initial object is the same as the verb (action) of the sentence; the predicates are included

[2] prefixes or suffixes

as properties. The method can also have some dependents. For example, the initial objects for the frames generated in the previous part are created as shown in figure 4.



* The verb "است" (is) represents the first person singular in Persian language.

Fig. 4. The initial objects obtained from sample sentences

4.1.4 Generation of final objects

Presence of references as pronouns and relations between sentences necessitates us to combine the objects. In this stage, several objects may form an object, i.e. all the properties and methods pertaining to a subject are merged into an object. Moreover, an object may be converted into several objects. For instance: two objects of the previous example are converted into the two objects displayed in figure 5.

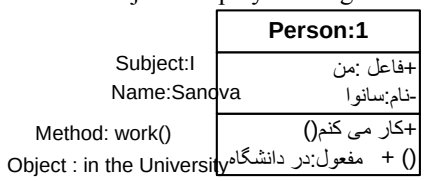


Fig. 5. The final objects obtained from sample sentences

According to former discussions, the lexicological relations between words are stored in the database. Variety of object-based relations might be generated in the objects combination stage; e.g. an object may inherit from another object.

5 Receiving the answers and evaluation

The users' answers are separately received in the form of natural language. They are converted again into objects for evaluation. For this purpose, the module explained in section 5 is applied. The objects obtained from users' answers will be compared with objects generated from correct answers allowing the users' grades to be evaluated.

For certain questions, the order of answer parts is important; therefore, at the time of generation of object from the correct answer, each part is assigned an identifier which determines the order. The sensitivity to answer is also activated for that question.

Sometimes mentioning all the items is required for gaining the whole grade in questions having answers in the form of lists of items, or the grade is divided so long as some of the items are responded. These facts are also stored when generating the objects from the correct answers for every question.

6 Persian language and investigation of system stages

Persian, also known as Farsi or Parsi, is an Indo-European language spoken and written primarily in Iran, Tajikistan and parts of Afghanistan. Persian alphabet contains 32 letters. Persian is written from right to left. Some other languages like Arabic, Kurdish, and Urdu use Persian's form of penmanship but have their own specifications. Persian also has its own specifications such as not using accents (except in special cases) and polymorphism in writing. The language has remained remarkably stable since the eighth century. It has a subject-object-verb word order, but has some head-initial structures [11].

Based on the dependency grammar, there are a number of dependents for any verb in every language indicating the capacity of that verb. For example, in [8, 12, 13, 7] the verb dependents have been determined for English, Chinese, German and Persian languages respectively. The current discussions are based upon the verb dependents in Persian language. Considering the free word order structure of Persian language, every sentence can be written in several forms while all the forms convey the same meaning. We generate the same object from different forms of a sentence with the aid of the objects generated by dependency grammar.

In the implementation for Persian language, a text is initially received by the user interface. The determination of the boundaries of words and sentences below the first phase sub modules, class tagging is performed for each word. The results of this stage are saved in a XML file. For example, if the answer for the question "What does Sanova do?" is the sentences "She goes to a journey. She travels with bus"; then, according to the correct answer, the wordclass.xml is firstly created, and the tagging operation is subsequently done in sentence level by performing syntactic processing on this file resulting in the creation of pos.xml file as indicated in figure 6.

```
<text>
  <sentence1>
    <subject>سانوا</subject>
    <object>به سفر</object>
    <verb>می رود</verb>
  </sentence1>
  <sentence2>
    <subject>او</subject>
    <object>با اتوبوس</object>
    <verb>مسافرت می کند</verb>
  </sentence2>
</text>
```

Fig. 6. The structure of POS.xml file

The frames of each sentence are generated using the XML files and then the corresponding objects of the frames are created. The created frames and objects have been illustrated in figure 7.

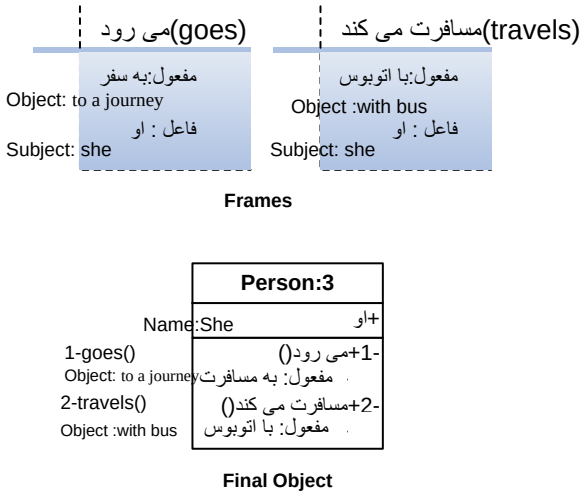


Fig. 7. The structure of the created frames and objects

When the set of objects was generated for each correct answer, the examinee inputs the relevant answer to each question via the user interface; the user's answer is in turn converted to a relevant object. For example, if in the answer to the previous question the user replies "Sanova goes to a journey"; following the creation of respective XML files, the object demonstrated in figure 8 is created.

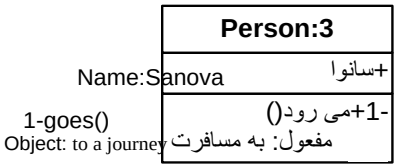


Fig. 8. The object generated from user's answer

The object generated from user's answer is compared with the object created from the correct answer. The two objects may not be completely identical but they might have common methods or attributes for which the user's grade must be evaluated based on the discussions in section 6.

7 CONCLUSION AND FUTURE WORK

Understanding the natural language by computers is one of the important matters in the area of information recovery systems. Many phrase structure algorithms have been proposed for natural language processing. The discrepancy of phrase structure system in the free word order languages revealed the significance of using dependency grammar algorithms.

Applying dependency grammar and conversion of natural language into objects, the computers can have a better comprehension of the natural language. The results of understanding the natural language can be used in different applications such as text summarization, text categorization, text translation systems and so on. We applied the system in the evaluation and grading system of essay questions in order to manifest the capability of the proposed method.

The complexity of natural language processing and the phrase structure method of the natural language processing have resulted in the absence of considerable researches in the field of devising machines for evaluation of users' answers to essay questions. Using

the Recommended method, a better comprehension of the natural language can be provided for the computers.

Acknowledgements. The author thanks Dr. Omid Tabibzadeh and Marzie Badiie for explaining all the tricky details of the Persian Dependency grammar.

References

[1] R. Canne, R. Kempson, and L. Marten, The Dynamics of Language: Elsevier Academic Press , 2005.

[2] D. Jurafsky, and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2 ed., Upper Saddle River, New Jersey: Pearson, Prentice Hall , 2009.

[3] N. Chomsky, The Logical Structure of Linguistic Theory: Plenum, 1956.

[4] M. Lester and L. Beason, The McGraw-hill handbook of english grammar and usage, 1ed, New York, London, Singapore: McGraw-Hill , 2005.

[5] L. Tesnière, Esquisse d'une Syntaxe structurale, Paris: Klincksieck, 1953.

[6] S. Kübler, R. McDonald, and J. Nivre, Dependency Parsing: Morgan & Claypool , 2009.

[7] O. TabibZadeh,Verb Valency and Basic Sentence Structures in Modern Persian (A Dependency-Based Approach),Tehran, Nashr-e Markaz Publishing Co , 2006.

[8] D. J. Allerton, Valency and the English Verb, London: Academic Press , 1982.

[9] U. Engel, Kurze Grammatik der deutschen Sparche, München: Iudicium Verlge , 2002.

[10] R. DesMarais, Student Success Handbook: New Readers Press, 2008.

[11] M. Iranpour Mobarakeh and B. Minaei-Bidgoli, "Verb Detection in Persian Corpus," International Journal of Digital Content Technology and its Applications, vol. 3, March 2009.

[12] W. Li. A Dependency Syntax of Contemporary Chinese. Institute of Linguistics,Chinese Academy of Social Sciences, manuscript, 1989.

[13] K. Fischer, German-English Verb Valency: A Contrastive Analysis. Tübingen: Gunther Narr, 1997 .