Ontology-Based Automatic Text Summarization Using FarsNet

Majid Ramezani[1], Mohammad-Reza Feizi-Derakhshi[2]

[1]*Department of Computer Engineering, University College of Nabi Akram,*
Tabriz, Iran
E-mail: sir.ramezani@gmail.com

[2]*Department of Computer Engineering, University of Tabriz,*
Tabriz, Iran
E-mail: mfeizi@tabrizu.ac.ir

Abstract

To summarize a text means to compress the text source into a shorter text in a way that the informational content is kept the same. With regard to the irregular volume of information available on the internet, manual summarization of huge volume of information by humans will be very arduous and difficult. There have been many activities in the field of automatic summarization so far. However, a lack of having methods capable of achieving a semantic hierarchy available in the documents is still felt. In this article we will propose a method for summarizing Persian documents which uses ontology for recognizing the semantic relationship between different parts of a text and extracting important sentences. For this purpose, mapping an input document with ontology brings about a graph whose vertices are the concepts available in ontology and its edges are the relationships among these concepts. This graph which is a graph-based representation of the input text comprises the necessary computational base for recognizing the important sentences in the input document and the production of a summary. The achieved results indicate the acceptable capability of the proposed method in obtaining semantic relationships available in documents and automatic text summarization. The FarsNet ontology is the base for this article.

Keywords: *Automatic text summarization, Ontology, FarsNet, Ontology-based automatic text summarization*

1. Introduction

“Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user or task” [1]. The growth and the irregular increase of the available information on the net or in other words the information overload [2], more than ever emphasizes the necessity of having a substitution method for presenting and choosing the textual contents. Based on such a method the most important parts of a text are said and by studying those parts, the reader can achieve the main ideas behind them or can make decisions about studying the texts thoroughly. Undoubtedly it is very difficult for humans to study and summarize a massive volume of textual documents available on the internet, by considering the time limit perhaps impossible.

Various classifications for automatic text summarization systems have been presented from different points of view ([3-6] for more study). Summarization Systems based on their output are grouped into *extractive* vs. *abstractive* and *generic* vs. *query-based*. In extractive text summarization systems the summary includes the most important parts available in the text (sentences, paragraphs, etc.), without making any changes [4], whereas for abstractive systems the summary is achieved by comprehension and then retelling the original text in fewer words [7]. Likewise generic text summarization systems are those in which the summary includes the most important parts available in the document while in query-based systems the summary only includes the concepts which are closely related to the query [8-9].

The summarization systems are classified into the following three groups according to the methodology used and the level of processing: *surface-level*, *entity-level*, and *discourse-level*. In surface-level systems the summary is brought about by depending on a collection of shallow features such as thematic features (presence of statistically salient terms, based on term frequency statistics), location (position in text, position in paragraph) and background (presence of terms from the title or headings in the text). In entity-level systems entities as well as the relation between them such as similarity, proximity, co-occurrence, co-reference and syntactic relations, are modeled by an internal representation for the text. In these systems generally for the sake of recognizing saliency, using structures such as graph topology, a representation of patterns of connectivity in the text is created. In discourse level systems the summary is also created with modeling the overall structure of the text and the relation available in it, in order to achieve communicative goals. The considered structures for the text in these systems can include the format of the document, threads of topics and rhetorical structure of the text.

According to another classification, automatic text summarization methods are generally divided to two groups: *supervised* methods and *unsupervised* methods [1]. Supervised methods usually use training sets of documents and associated summaries, which based on that the original sentences are labeled as relevant or non-relevant, for inclusion in the summary [4; 10-12]. After training, by working on unlabeled texts the system rank the sentences due to their relevance in order to include them in the summary. On the other hand, unsupervised methods by depending on a set of rules and without any need to training sets, engage in summarizing documents [8; 13-14]. Unsupervised methods can be categorized into the following three groups: *vector-space* method, *graph-based* method and *text-based* methods, according to the representation performed from the input text and their work space. In methods based on vector space, after representing text units (terms, sentences, paragraphs or documents) in a format of vectors in a vector space, the most relevant units are chosen to be present in the summary, by using techniques like Latent Semantic Analysis (LSA) [15-16] or Maximal Marginal Relevance (MMR) [17], (e.g. [18-20]). In graph-based methods the goals of summarization are followed by representation of each of the text units in the format of graphs' vertices and their relations in the form of graphs' edges [21], (e.g. [22-24]). Finally, in text-based methods the processed text is summarized without being transferred to another work space [25].

In order to make an intermediate representation of input text in ontology-based automatic text summarization, we use the ontology knowledge-base. Ontology which is a philosophical concept, is the explicit specification of a conceptualization [26]. Ontology is a word with a Greek etymology which is comprised of the two words; "onto" which means "beings" and "logos" which generally means "science". Thus it can be said ontology is the science or study of beings [27]. Based on interpretation ontology is a systematic account of existence [26], which is used for modeling the real world's entities and the relation among them. In knowledge-based systems everything that exists, is exactly what its representation is possible [26]. It can be inferred that in different tasks of the artificial intelligence, what can exists in a conceptualized world, is determined by what can be represented [28]. Therefore, we will use the ontology for representing the present world for the machine, so that by using the summarization methods which have the ability to extract semantic hierarchy among the entities, we specifically pursue the goals of automatic text summarization. Today there is also a need to pay attention to semantics in Information Extraction (IE) systems, in order to improve the usage of the available information and this has been emphasized [29-30].

The aim of this paper is to design and implement an extractive, generic, entity-level, unsupervised and graph-based text summarization system, for Persian documents for which ontology is the basis for forming summaries. In this framework in section two, we will consider the works done in this field and then in section three the ontology of FarsNet will be introduced. In section four our attention will go to the architecture of the proposed system. In section five by considering the method of evaluating the automatic text summarization systems, the results of the proposed system will be analyzed. Finally the sixth section is dedicated to conclusion.

2. Related Works

The first automatic text summarization system was presented in 1958 [29], in which the summary is produced by using a set of shallow features. By the passage of time, there have been many efforts in this field (e.g. [18; 25; 31-40]), it is in a way that we can say this branch has reached its maturity today [41]. However, the background for using ontology for the purpose of automatic summarization and the improvement of its achieved results does not have an old history.

Wu and Liu [42] have dealt with the comparison of the two methods of summarization; one based on term frequency and another based on ontology in business news articles. They have recognized the most important topics by using ontology and they have determined the relative significance of each of them as well. Then the paragraphs will be chosen with regard to the size of the summary by ranking each of the paragraphs according to their relevance to the title of the document. In [21] for the sake of the improvement of semantic representation of the source text, the ontology is used as specific domain knowledge for biomedical extractive text summarization. Hennig et al. [43] achieved a better representation of sentence's information content by matching the sentences of the text with the ontology's vertices. Using a classifier, they have chosen the best sentences for presenting in the summary, by depending on the extracted features from the ontology for each sentence.

Viswanath [44] in order to present a method beyond statistical methods, has proposed a knowledge-based method which has used ontological knowledge for determining the importance of the sentences. The ontology used in this research is Wikipedia ontology. In [45] the ontology for multi-document summarization (in which the summary includes the most important notions from more than one document) is also used. Bawakid [46] in his doctoral thesis by using WordNet [47] has dealt with the semantic similarity among the terms. Then by quantifying these similarities, with regard to their semantic content, the sentences will be scored and finally the sentences with the highest scores will be chosen for inclusion in the summary. In this study a module is also used to reduce redundancy. Zhu et al [48] have proposed a method for summarizing the web pages based on ontology, in which the importance of each of the sentences have been evaluated by considering a combination of topic concepts and webpage structure and then the summary is formed. Yago-based Summarizer is another sample of these systems which was presented by Baralis et al [49]. In this system in order to achieving the real meaning of the texts, some steps like entity recognition and disambiguation based on Yago ontology [50] for automatic text summarization have been considered. [51] By doing a review on the recent progress in the field of ontology-based summarization, with a dependence on methods such as socio-cognitive and structural discourse models, has been introduced the Textminer as an automatic text summarization system. Li and Li [52] have suggested a system for summarizing documents in the field of disaster management, by studying the possibility of using ontology in multi-document summarization.

With regard to the fact that in all the mentioned studies, ontology is the basis for forming the summaries what is common among them as expected, is the level of process of entity-level in them. All of the works pursue to obtain semantic levels available in the text in summarization goals by using ontology. There have been no studies performed in the field of summarizing Persian documents based on ontology so far. The system presented in this study with acting in the entity level will deal with summarizing documents by using the FarsNet ontology.

3. FarsNet

WordNet [47] is a large electronic lexical database of English which has been designed by Princeton University and includes nouns, adjectives, verbs and adverbs which are classified into the sets of cognitive synonyms (synset). There have been several versions of WordNet designed for different languages like German, French, Spanish, Dutch, Italian and etc. so far. FarsNet is also a version of WordNet which has been devised by the laboratory of natural language processing of the Shahid Beheshti University for Persian language. The first version of the Persian WordNet [53] includes word, syntactic and semantic knowledge for more than 15000 Persian words and phrases, and 1000 synsets which is formed by nouns, adjectives and verbs. Likewise WordNet, FarsNet also organizes words in collections which are cognitively synonymous and these collections are called synonym set or synset. There is a noticeable difference in the size of this lexical database and that of the WordNet of the University of Princeton which includes 155,287 word entries and 117,659 synsets. The most important relation between the words in FarsNet is the synonymy relation. FarsNet is designed in a way which its connectivity with WordNet 3.0 is possible and the relation between equal-to and near equal-to among the synsets exists.

The second version of this lexical database [54] consists of more than 30,000 lexical entries from parts of speech like nouns, verbs, adjectives and adverbs and more than 20,000 synsets. The relations between senses and synsets in FarsNet are like the relations available in WordNet. Relations such as anthonymy, synonymy, hypernymy, hyponymy, meronymy.

4. Automatic Summarization System

The proposed summarization system in this study is made of two phases: *preprocessing* which includes all the necessary activities in order to extract information and recognize the lexical relationships available in the text and the *processing* phase in which regarding the information extracted in the previous phase, the necessary processes are performed to select the most suitable sentences and produce the summary. In what follows we will study the details related to each of these phases.

4.1. Preprocessing

This phase whose aim is to create the necessary calculations to produce the summary in the next phase, itself includes several steps. These steps respectively are:

A. Tokenization: In this level as the first level of the system architecture, the input document as a stream of text is decomposed to comprising tokens (words, symbols and other meaningful elements).

B. Sentence Boundary Detection: This level includes the recognition of the boundary of sentences as the desired text unit. For this purpose Persian punctuation marks including (".", "!", "?") which are used at the end of sentences, have been used.

C. Anaphora Resolution: Anaphora describes the language phenomenon of referring to an entity (an object or event) which was mentioned before and anaphora resolution is the process of finding these entities [55]. Consider a document whose subject is diabetics. The name of this disease may appear only once in the beginning of the text and after that, pronouns and referral phrases are used. By considering the sentences as the text units, these references will be definitely disconnected. Moreover there will be problems when each sentences' elements are matched with ontology entries, and consequently the authenticity of the results will be reduced. Therefore it seems necessary to replace pronouns and referral phrases with the corresponding references in each high-level task from the natural language processing, in order to maintain the semantic structure of sentences. In Persian, a pronoun as a grammatical word only refers to person and number and falls into seven types: *personal*, *adjoining personal*, *demonstrative*, *reflexive/emphatic*, *reciprocal*, *question* and *indefinite* pronouns. Considering the fact that a ready component for anaphora resolution in Persian is not available, we have just attempted to replace personal and demonstrative pronouns with the corresponding references and we abandoned to replace other types of pronouns as well as referral phrases with their references. Consider the following example:

[Grahame Bell] is the [first inventor of practical telephone]. [He] was one of the founding members of the National Geographic Society. [The inventor of telephone], [who] was born in Edinburgh, Scotland, was educated at the University of London.

By identifying the personal pronouns (I, you, he, she, it, you, we, and they) as well as the names of people in Persian language, the recognized pronouns will be replaced with the first name which has been previously mentioned in the previous sentence. To recognize the names of people, we have used the Persian version of Wikipedia. In the above example [He] replaces with [Graham Bell]. For adjoining pronouns which accompany another word such as “کتابم” (my book), “کتابت” (your book), “کتابش” (his/her book), “کتابمان” (our book), “کتابتان” (your book) and “کتابشان” (their book), based on the stemming conducted in the next step, the adjoining pronoun is removed and only the stem remains (“کتاب”, book). For demonstrative pronouns “این” (this) and “آن” (that) a replacement with the first noun before it (the nearest noun) takes place. For the rest of remaining pronouns including reflexive/emphatic (myself, himself), question pronouns (who, what, which), indefinite pronouns (all, no) and reciprocal pronouns (each other, together) no replacement has taken place.

In this article no attempts have been made to replace the referral phrases with the corresponding references which regarding to the ontology-based approach, is justifiable. Namely, in the above example, [Grahame Bell] who is an instance of “inventor” (and more specifically he is “the inventor of telephone”), is semantically related to the phrase [the inventor of telephone] and their entities in the graph representation of the input text (which will be produced in the last step in the preprocessing phase), will be connected and adjacent.

Undoubtedly, the approach used in this paper will have some shortcomings, but as mentioned because there is no component available for this purpose, and the necessity of performing anaphora resolution, this primary method has been used. Also as the emphasis of this research is on ontology-based automatic summarization, this approach has been considered as sufficient. The achieved results imply the improvement of summarization process when using this primary approach in comparison to the time it is not used. Obviously the application of comprehensive approaches and with a high reliability for anaphora resolution will result in better results of summarization.

D. Stemming: Stemming is the act of reducing the words that are morphologically similar to each other into a simple term which is called stem or root [56]. For example the word “مشکلات” (difficulties) reduce to the stem “مشکل” (difficulty). Regarding the fact that the FarsNet knowledgebase is made of the stems relating to the entities not their different morphological forms, it seems necessary to achieve the stems of words to match them with the ontology. To fulfill this aim we have used the automatic stemmer of Persian words presented by Nojavan, Ramezani and Feizi-Derakhshi [57]. This automatic stemmer uses a combination of Persian lexical rules and a database in order to get the stems.

E. Named Entity Recognition (NER): NER is the recognition of the important names in the text like the names of people, organizations and locations [58]. In this study, to recognize named entities automatically, based on the fact that the ontology is itself the hierarchical database for entities, all the available words in input text, have been searched in FarsNet and if there is a match in each search, a named entity will be recognized. The FarsNet 2.0 is the base for ontology-based calculations in this study. It is evident that the authenticity resulted from this section depends on the number of available entities used in the ontology. The recognized entities comprise the graphs' nods which are the conceptual base for selecting sentences for inclusion in the summary. By extracting the relations between the recognized entities, the graph will be completed in the next step.

F. Relation Extraction: As mentioned before the work space in this study is graph-based. The goal of this level as the last level in the preprocessing phase is the completion of the graph topology and creating a graph-based representation of the input text. After the entities have been recognized in the previous stage, in this stage by considering the ontology, the semantic relations among each of these entities will be recognized and these relations as edges between vertices (entities) of graph will be drawn. To do this, we will use a collection of semantic relations including hypernymy/hyponymy, meronymy, antonymy, and synonymy, which has been extracted from FarsNet. If the vertex (entity) ν_i happens in the synset of the vertex ν_j a directed edge will be drown between them (from ν_i to ν_j). At the same time, synsets are related via semantic relations such as hypernymy, hyponymy and meronymy. If X is a subtype or an instance of Y, then X is a hyponym of word Y (or Y is a hypernym of word X). For instance “Graham Bell” is a hyponym for “inventor” (as well as “inventor” a hypernym for “Graham Bell”). If X is a part of Y, then X is a meronymy of Y. For example, finger is a meronymy of hand. These relations exist among synsets. If the relations of Hypernymy/Hyponymy, Meronymy exist between corresponding synsets of each vertices a directed edge will be drawn between them. Also by extracting the antonymy relation between entities from FarsNet, provided that there is an antonymy relation between two vertices, a directed edge will be drawn between them. The idea in this approach is that the existence of any of the mentioned relationships among the vertices of the graph, represents the semantic relation between corresponding entities. By doing this a graph-based representation of the input text is created which includes the semantic relations among their entities and contains a semantic schema of the input document. The resulting graph is a small sub graph of the set of the entities available in the ontology and the relations among them. We should pay attention that since it is possible to have more than one relation from the mentioned ones between two vertices, it is also possible to draw more than one edge from one vertex to another in the graph (directed multigraph).

Figure 1 represents an illustration of the available steps in the first phase of the summarization system. As it is seen the input text, is the input of the phase and its output is a graph representation of the input text.

4.2. Processing

In this phase the gained information from the previous phase will be processed to produce a summary in two steps. These two steps are:

A. Graph Analyzer: as mentioned, in extractive summarization, the summary is obtained by choosing a subset of sentences from the original document. For this purpose a set of the most central sentences containing the most important information available in the original document, is chosen. The centrality of a sentence is usually determined by the centrality of its words [59]. The aim of this section is to evaluate each of the entities to be able to decide the centrality of each of the sentences in the next stage, in order to achieve a sort of ranking. In this research we have used three evaluation measures which include *Degree centrality* [60], *Eigenvector centrality* [61] and *Barycenter centrality* [62] to evaluate the centrality of each of vertices of graph (entities).

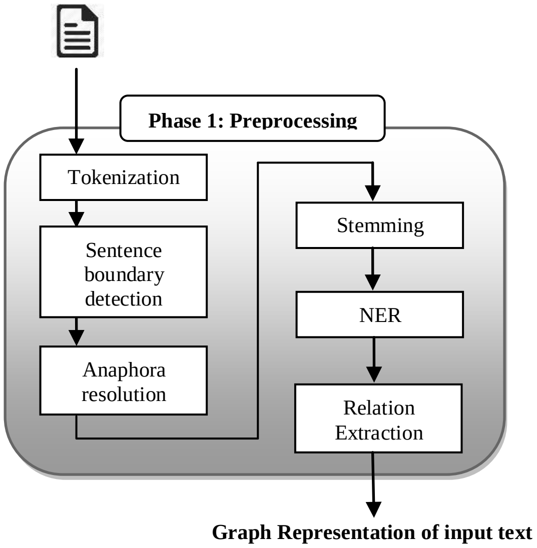

Fig. 1 Preprocessing Phase

● **Degree Centrality:** Based on this measure the centrality of each vertex in the graph equals the number of its relations or edges. In other words the centrality of each vertex equals its degree. With regard to the fact that the graph is directed, the degree of each vertex equals the total number of input and output edges to/from it. The idea behind this measure is that a high degree of each vertex in the graph represents more relations to the other vertices and consequently its higher semantic importance. Therefore it can be said in brief that in a graph the degree centrality of a vertex, equals its relations with other vertices of the graph.

● **Eigenvector Centrality:** This measure is an expansion of the degree centrality measure, with this explanation that unlike the degree centrality measure, the centrality of a vertex is not only dependent on its degree, but the degree of the vertices which are in relation with the desired vertex, will influence the centrality. In other words according to this measure, the centrality of a vertex will be high in case it is in relation with other vertices with high degrees. Thus a vertex with a few relations with high degree vertices will have a higher centrality in comparison with a vertex with more relations with low degree vertices. Therefore the eigenvector centrality of each vertex equals the degree of it in addition to the total degrees of the vertices which have a direct relation (a direct edge) with it.

● **Barycenter Centrality:** Unlike the two previous measures which the centrality of the vertices depend on their degree or the degree of adjacent vertices, in this measure total distance (the number of edges in the shortest path) from each vertex to all other vertices, represents its centrality. According to this measure, barycenter centrality of vertex ν equals 1/ total distance from vertex ν to all other vertices. If the total distance from vertex to all other vertices is high, the vertex has a lower centrality, since it doesn't have a direct (semantic) relation with other vertices, and vice versa. The highest value of this centrality is obtained when the distance from current vertex to all other vertices equals one. It means that there is a direct edge between the current vertex and any other vertices and in such case the corresponding entity definitely has a high significance, which has a direct semantic relation with other entities. If there is no path between the two vertices the distance between them will be considered as indefinite.

B. Summary Production (sentence scoring, ranking and selection): the aim of this stage as the final stage in the system architecture is to determine the centrality of each of the sentences of the original document and ultimately choosing the most central of them for inclusion in the summary. For this purpose at first after determining the centrality of each entity in the previous step, we calculate the centrality of each sentence as its *centrality score*. The centrality score of each sentence equals the sum of centrality of all the entities available in it which according to one of the mentioned measures in the previous stage have been evaluated. We should pay attention in order to avoid sacrificing shorter sentences for longer sentences, we should normalize this score by the length of the sentences. Then all the sentences based on their centrality scores are ranked and at the end the sentences with the highest scores, regarding to the *compression rate* which represent the size of the summary than to the original document, will be chosen for inclusion in the summary. We have used a 30% compression rate for all the summarization processes during this study which is a common compression rate.
Figure 2 represents an illustration of the available steps in the second phase of the summarization system. As it is seen the input of this phase is a graph representation of the input text and its output is the summary.

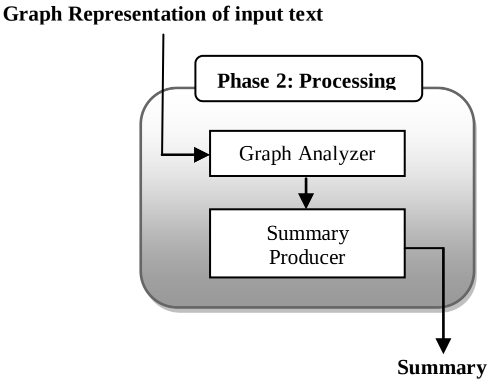

Fig. 2 Processing phase

5. Evaluation Experiments and Results

In this study a system was devised for the purpose of automatic Persian text summarization whose functionality base is FarsNet ontology. In what follows we deal with the evaluation of the system.

5.1. Evaluation Condition

The evaluation methods of the automatic summarization systems generally are divided into two main sections: *extrinsic* and *intrinsic* methods [63]. In extrinsic evaluation methods the quality of the summaries in performing certain tasks (like Information Retrieval (IR)) is evaluated, while in intrinsic methods the summaries independently and based on the analyses from the summaries are evaluated ([63] is suggested for more study).

The base of comparison for the automatic summarization systems is the summaries which have been produced by humans and are called *golden* or *reference summary*. The golden summaries for a set of documents are being produced by human here.

We have used intrinsic evaluation measures including *precision*, *recall* and also *F-score* measure (as the combination of precision and recall) to evaluate the obtained results of automatic summarization. The precision is the fraction of retrieved instances that are relevant and the recall is the fraction of relevant instances that are retrieved [64]. With regard to the fact that we deal with sentences as desired text units, we can express that the precision equals the number of common sentences between the golden summary and *system summary* (the summary which is produced by system), divided by the number of the sentences of the system summary. The recall equals the number of the common sentences between the golden summary and the system summary, divided by the number of the sentences of the golden summary [63]. The F-score measure equals the harmonic average of the precision and recall measures and equals (2×P×R)/(P+R).

5.2. Evaluation Results

As mentioned before, in this study there are three ways to evaluate the centrality of the graphs vertices and consequently the centrality of sentences, including the measures of degree centrality, eigenvector centrality and barycenter centrality. Thus there are three possible methods to produce automatic summary. Table 1 includes the evaluation results of the summaries produced by using each of the three methods for precision, recall and F-score measures. It should be mentioned that values presented in this table are the results of average values, obtained from evaluation of a set of documents.

Table 1. Evaluation results for ontology-based summarization

Centrality Evaluation Metric	**Precision** (%)	**Recall** (%)	**F-Score** (%)
Degree Centrality	61.34	58.62	60.24
Eigenvector Centrality	**65.40**	**61.82**	**63.92**
Barycenter Centrality	58.22	54.76	56.84

As it is seen the eigenvector centrality has got the best results among these three measures. In other words, among the three measures of centrality the measure which considers the degree of each vertex in addition to the vertices related to it, will have a better function in the evaluation the importance of different textual texts. The summaries produced by using degree centrality measure achieved higher values in comparison to barycenter centrality measure for the precision, recall and F-score. It can be deduced that in evaluating the centrality of the vertices of the graph, the measures which are based on the degrees of vertices (in the first rank is the one which contains the degree of the current vertex in addition to the adjacent vertices and the second rank is the measure which only contains the degree of current vertex), will have a better function in comparison to those which are based on distance of vertices from each other. In spite of that the obtained results show an acceptable quality in produced summaries, we should pay attention to the fact that the FarsNet has been the base for decision making for sentence selection and summary production, so it is expected that using the future and more complete versions of that, which contains more entities and more semantic relations, leads to better results in automatic summarization.

6. Conclusion

The existence of an automatic text summarization system will definitely facilitate it for one who deals with reading and results in time budgeting. In this paper with the aim of ontology-based automatic summarization of Persian documents, a system has been proposed which recognizes the semantic relations available in the text by using the FarsNet ontology and extracts the most important sentences for inclusion in the summary. The obtained results from the evaluation of the produced summaries indicate the acceptable success of the proposed method in using the ontology of FarsNet in summarizing Persian documents. It can be inferred that ontology has an effective role in modeling and conceptualizing the real world for machine in the task of automatic text summarization.

References

[1] I. Mani and M. T. Maybury, *Advances in automatic text summarization*. the MIT Press, 1999.

[2] K. Ježek and J. Steinberger, “Automatic Text Summarization (The state of the art 2007 and new challenges),” in *Proceedings of Znalosti*, 2008, pp. 1–12.

[3] E. Lloret, “Text summarization: an overview,” *Pap. Support. by spanish Gov. under Proj. TEXT-MESS*, 2008.

[4] A. Nenkova and K. McKeown, “A survey of text summarization techniques,” in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer, 2012, pp. 43–76.

[5] E. Lloret and M. Palomar, “Text summarisation in progress: a literature review,” *Artif. Intell. Rev.*, vol. 37, no. 1, pp. 1–41, 2012.

[6] H. Saggion and T. Poibeau, “Automatic text summarization: Past, present and future,” in *Multi-source, Multilingual Information Extraction and Summarization*, Springer, 2013, pp. 3–21.

[7] V. Gupta and G. S. Lehal, “A survey of text summarization extractive techniques,” *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, Aug. 2010.

[8] J.-H. Lee, S. Park, C.-M. Ahn, and D. Kim, “Automatic generic document summarization based on non-negative matrix factorization,” *Inf. Process. Manag.*, vol. 45, no. 1, pp. 20–34, Jan. 2009.

[9] I. Mani, M. T. Maybury, and M. Sanderson, *Advances in automatic text summarization*, vol. 26, no. 2. the MIT Press, 1999, pp. 280–281.

[10] Y. Chali and S. a. Hasan, “Query-focused multi-document summarization: automatic data annotations and supervised learning approaches,” *Nat. Lang. Eng.*, vol. 18, no. 01, pp. 109–145, Apr. 2012.

[11] M. Litvak and M. Last, “Multilingual Single-Document Summarization with MUSE,” *MultiLing 2013*, p. 77, 2013.

[12] C. Li, X. Qian, and Y. Liu, “Using Supervised Bigram-based ILP for Extractive Summarization.,” in *ACL (1)*, 2013, pp. 1004–1013.

[13] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. a. Mehdiyev, “MCMR: Maximum coverage and

minimum redundant text summarization model," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14514–14522, Nov. 2011.

[14] O. Gross, A. Doucet, and H. Toivonen, "Document Summarization Based on Word Associations," in *Proceedings of the 37th international ACM SIGIR conference on Research and Development in Information Retrieval. ACM*, 2014.

[15] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," *Psychol. Rev.*, vol. 104, no. 2, p. 211, 1997.

[16] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, *Handbook of latent semantic analysis*. Psychology Press, 2013.

[17] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335–336.

[18] D. Ai, Y. Zheng, and D. Zhang, "Automatic text summarization based on latent semantic indexing," *Artif. Life Robot.*, vol. 15, no. 1, pp. 25–29, 2010.

[19] M. M. G. Ozsoy, F. F. N. Alpaslan, and I. Cicekli, "Text summarization using latent semantic analysis," *J. Inf. Sci.*, vol. 37, no. 4, pp. 405–417, 2011.

[20] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, no. 2, pp. 4985–4988.

[21] L. P. Morales, A. D. Esteban, P. Gervás, I. Matveeva, C. Biemann, M. Choudhury, and M. Diab, "Concept-graph based biomedical automatic summarization using ontologies," in *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, 2008, no. August, pp. 53–56.

[22] S. Miranda-Jiménez, A. Gelbukh, and G. Sidorov, "Summarizing Conceptual Graphs for Automatic Summarization Task," in *Conceptual Structures for STEM Research and Education*, Springer, 2013, pp. 245–253.

[23] R. Ferreira, F. Freitas, L. De Souza Cabral, R. Dueire Lins, R. Lima, G. Franca, S. J. Simskez, and L. Favaro, "A Four Dimension Graph Model for Automatic Text Summarization," *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1. pp. 389–396, 2013.

[24] Y. Ledeneva, R. A. García-Hernández, and A. Gelbukh, "Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2014, pp. 466–480.

[25] H. P. Edmundson, "New methods in automatic extracting," *J. ACM*, vol. 16, no. 2, pp. 264–285, 1969.

[26] T. T. R. Gruber, "What is an Ontology," 1993.

[27] T. Lawson, "A conception of ontology," *Mimeogr. Univ. Cambridge*, 2004.

[28] T. R. Gruber and G. R. Olsen, "An Ontology for Engineering Mathematics.," *KR*, vol. 94, pp. 258–269, 1994.

[29] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.

[30] S. Ranwez, B. Duthil, M. F. Sy, J. Montmain, P. Augereau, and V. Ranwez, "How ontology based information retrieval systems may benefit from lexical text analysis," in *New Trends of Research in Ontologies and Lexical Resources*, P. Vossen, L. Qin, and E. Hovy, Eds. Springer, 2013, pp. 209–231.

[31] J. J. Pollock and A. Zamora, "Automatic abstracting research at chemical abstracts service," *J. Chem. Inf. Comput. Sci.*, vol. 15, no. 4, pp. 226–232, 1975.

[32] R. Brandow, K. Mitze, and L. F. Rau, "Automatic condensation of electronic publications by sentence selection," *Inf. Process. Manag.*, vol. 31, no. 5, pp. 675–685, 1995.

[33] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Advances in automatic text summarization*, I. Mani and M. T. Maybury, Eds. MIT Press, 1999, pp. 111–121.

[34] K. Zechner, "Automatic summarization of spoken dialogues in unrestricted domains," no. November, 2001.

[35] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, and D. Liu, "MEAD-a platform for multidocument multilingual text summarization," 2004.

[36] A. Nenkova, "Automatic text summarization of newswire: Lessons learned from the document understanding conference," in *AAAI*, 2005, vol. 5, pp. 1436–1441.

[37] E. Lloret, O. Ferrández, R. Munoz, and M. Palomar, "A Text Summarization Approach under the Influence of Textual Entailment.," in *NLPCS*, 2008, pp. 22–31.

[38] H. J. Jain, M. S. Bewoor, and S. H. Patil, "Context Sensitive Text Summarization Using K Means Clustering Algorithm," *Int. J. Soft Comput. Eng.*, vol. 2, no. 2, pp. 301–304, 2012.

[39] V. Gupta and G. S. Lehal, "Automatic Text Summarization System for Punjabi Language," *J. Emerg. Technol. Web Intell.*, vol. 5, no. 3, pp. 257–271, 2013.

[40] V. Qazvinian and D. Radev, "Generating extractive summaries of scientific paradigms," *J. Artif. Intell. Res.*, vol. 46, pp. 165–201, 2013.

[41] P. Dehkordi and F. Kyoumarsi, "Using gene expression programming in automatic text summarization," *Middle East J Sci Res*, vol. 13, no. 8, pp. 1070–1086, 2013.

[42] C.-W. Wu and C.-L. Liu, "Ontology-based Text Summarization for Business News Articles.," *Comput. Their Appl.*, vol. 2003, pp. 389–392, 2003.

[43] L. Hennig, W. Umbrath, and R. Wetzker, "An ontology-based approach to text summarization," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, 2008, vol. 3, pp. 291–294.

[44] M. Viswanath, "Ontology-based automatic text summarization," University of Georgia, 2009.

[45] L. Li, D. Wang, C. Shen, and T. Li, "Ontology-enriched multi-document summarization in disaster management," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 819–820.

[46] A. Bawakid, “Automatic documents summarization using ontology based methodologies,” University of Birmingham, 2011.

[47] G. Miller and C. Fellbaum, “Wordnet: An electronic lexical database.” MIT Press Cambridge, 1998.

[48] J. Zhu, Y. Jiang, B. Li, and M. Sun, “Ontology-based Automatic Summarization of Web Document,” *Int. J. Adv. Comput. Technol.*, vol. 4, no. 14, pp. 298–306, Aug. 2012.

[49] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, and S. Shah, “Multi-document summarization based on the Yago ontology,” *Expert Syst. Appl.*, vol. 40, no. 17, pp. 6976–6984, 2013.

[50] F. M. Suchanek, G. Kasneci, and G. Weikum, “Yago: A large ontology from wikipedia and wordnet,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 6, no. 3, pp. 203–217, 2008.

[51] P. Hípola, J. A. Senso, A. Leiva-Mederos, and S. Domínguez-Velasco, “Ontology-based text summarization. The case of Texminer,” *Libr. Hi Tech*, vol. 32, no. 2, pp. 229–248, Jun. 2014.

[52] L. Li and T. Li, “An empirical study of ontology-based multi-document summarization in disaster management,” *Syst. Man, Cybern. Syst. IEEE Trans.*, vol. 44, no. 2, pp. 162–171, 2014.

[53] M. Shamsfard, “Developing FarsNet: A lexical ontology for Persian,” in *4th Global WordNet Conference, Szeged, Hungary*, 2008.

[54] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi, “Semi automatic development of farsnet; the persian wordnet,” in *Proceedings of 5th Global WordNet Conference, Mumbai, India*, 2010.

[55] R. Mitkov, *Anaphora resolution*. Routledge, 2014.

[56] V. Gupta, “Automatic Stemming of Words for Punjabi Language,” in *Advances in Signal Processing and Intelligent Recognition Systems*, Springer, 2014, pp. 73–84.

[57] M.-B. Nojavan, M. Ramezani, and M.-R. Feizi-Derakhshi, “Automatic Stemming of Persian Words Using An Optimal Combination of Lexical Rules and Database,” in *8th International Conference of Iranian Society for Promotion of Persian Language and Literature*, 2013, pp. 1–11.

[58] B. Mohit, “Named Entity Recognition,” in *Natural Language Processing of Semitic Languages*, I. Zitouni, Ed. Springer, 2014, pp. 221–245.

[59] G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” *J. Artif. Intell. Res.(JAIR)*, vol. 22, no. 1, pp. 457–479, 2004.

[60] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Soc. Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[61] P. Bonacich, “Some unique properties of eigenvector centrality,” *Soc. Networks*, vol. 29, no. 4, pp. 555–564, 2007.

[62] A. Beveridge, “Centers for random walks on trees,” *SIAM J. Discret. Math.*, vol. 23, no. 1, pp. 300–318, 2009.

[63] J. Steinberger and K. Ježek, “Evaluation measures for text summarization,” *Comput. Informatics*, vol. 28, no. 2, pp. 251–275, 2012.

[64] S. Rüger, “Multimedia information retrieval,” *Synth. Lect. Inf. Concepts, Retrieval, Serv.*, vol. 1, no. 1, pp. 1–171, 2009.

Majid Ramezani received his B.S. in Software Engineering from the Payame-Noor University. Then he received his M. Sc. in Artificial Intelligence from the University College of Nabi Akram. His research interests include: natural language processing, computational linguistics, psycholinguistics and neurolinguistics.

Mohammad-Reza Feizi-Derakhshi received his B.S. in Software Engineering from the University of Isfahan. Then he received his M.S. and Ph.D. in Artificial Intelligence from the Iran University of Science and Technology. He is currently a faculty member at the University of Tabriz. His research interests include: natural language processing, optimization algorithms, intelligent methods for fault detection and intelligent databases.