A Fuzzy Approach for Ambiguity Reduction in Text Similarity Estimation (Case Study: Persian Web Contents)

Hamid Ahangarbahan
Department of Information Technology Engineering, Tarbiat Modares, Tehran, Iran
ahangarbahan@gmail.com
Golam Ali Montazer*
Department of Information Technology Engineering, Tarbiat Modares, Tehran, Iran
montazer@modares.ac.ir

Received: 24/Jun/2015 Revised: 24/Nov/2015 Accepted: 05/Dec/2015

Abstract

Finding similar web contents have great efficiency in academic community and software systems. There are many methods and metrics in literature to measure the extent of text similarity among various documents and some its application especially in plagiarism detection systems. However, most of them do not take ambiguity inherent in word or text pair's comparison that gained form linguistic experts as well as structural features into account. As a result, pervious methods did not have enough accuracy to deal vague information. So using structural features and considering ambiguity inherent word improve the identification of similar contents. In this paper, a new method has been proposed that taking lexical and structural features in text similarity measures into consideration. After preprocessing and removing stop words, each text was divided into general words and domain-specific knowledge words. For each part, appropriate features and measures are extracted. Then, the two lexical and structural fuzzy inference systems were designed to assess lexical and structural text similarity respectively. The proposed method has been evaluated on Persian paper abstracts of International Conference on e-Learning and e-Teaching (ICELET) Corpus. The results shows that the proposed method can achieve a rate of 75% in terms of precision and can detect 81% of the similar cases.

Keywords: Text Similarity; Similarity Metric; Fuzzy Sets; Lexical Similarity; Structural Similarity; Persian Text.

1. Introduction

At present, a vast amount of text resources can easily be accessed on the Internet. Although such high frequency of web documents provides us with rapid and immediate access to information, similar and duplicated data results in waste of time and confusion on the part of researchers who would like to detect the originality of a document. Finding the similar contents recently attracts a lot of researchers. Text or content similarity detection can be employed in paraphrasing identification, plagiarism, text summarizing, sentiment analyses, text clustering, text entailment, tracking text news flow on web, etc. For instance, accurate text similarity detection leads to better performance in paraphrasing identification and can improve text clustering [1].

Numerous studies have been conducted to detect similar documents as well as plagiarism [2-3], but most of them have not taken the types of content and domain-specific knowledge as well as style and structural of writing into account. From another perspective, in these studies the methods of text similarity measurement which exert a major influence on the accuracy of evaluation have been used for a certainty and express in crisp way. For instance the word "Process" has different importance in image processing content versus e-learning. As a result, while comparing two texts in specialized or academic domain we will face ambiguity in word or text pair's comparisons since pervious methods don't consider structural features such as author's style of writing. Such limitations in similarity measurement reduce the effectiveness of previous methods in surface and semantic level of text [4-5]. To overcome these limitations, we have proposed a new method that can deal with the ambiguity in the similarity measurement and also consider structural features of text. This method deploys fuzzy linguistic variables to express experts' knowledge about text similarity in surface level of text. Two lexical and structural fuzzy inference systems were designed to accurately figures out the lexical and structural similarity respectively. The output of these two fuzzy inference systems are combined with specific factor and finally the extent of similarity between two contents is determined.

The rest of this paper is organized as follows: the following section provides a brief review of the most related studies. In Section 3text or content similarity problem has been described. Fundamental concepts of fuzzy set theory will be represented in Section 4, the architecture of the proposed method and numerical results has been provided in Section 5. Finally, Section 6 offers the conclusions and implications of the study.

* Corresponding Author

2. Related Work

This section provides the most related research in text similarity methods. Alzahrani, et al., did a comprehensive overview of all the text similarity methods proposed for plagiarism detection. They classified these methods in two broad categories: Literal and Intelligent. With respect to literal methods, individual use simple operations such as copy and paste, which is the most common form of plagiarism and in intelligent methods they use more sophisticated methods such as paraphrasing, obfuscation, changing the structure to hide cheating (redrafting), translation from one language to another and adopting other people's ideas. They also presented taxonomy of techniques that can be utilized to detect text similarity. That taxonomy was categorized into six groups: character-based, vector- based, grammar-based, semantics-based, fuzzy-based, and structure-based. They concluded and proposed semantic- and fuzzy- based methods to be applied due to their better detection and estimation of text similarity and plagiarism detection [8].

Osman, et al., also made another research on text similarity methods. Their proposed taxonomy almost resembled pervious work and consisted of six categories. They consider cluster-based and cross language-based instead of fuzzy-based and vector-based methods. They also proposed semantic role labeling for sentence to detect similarity. The main drawback of their methods is related to numerous computations needed to make [9].

Alzahrani and Salimi used fuzzy membership function to calculate the degree of similarity between two words. They obtained an accuracy of 54.24% on PAN-PC-09 corpus. Their algorithm suffers from time complexity and also needs to improve membership function [10]. In another study, Gupta, et al., drawing on the model offered by Alzahrani and Salimi, redefined fuzzy membership function in smaller range intervals. They also used different preprocessing on PAN-PC-2012 dataset and applied it to fuzzy algorithm. The authors concluded that preprocessing on POS level and its integration with fuzzy-based methods would contribute to effective recognition similar documents [11].

Based on their study, El-Alfy, et al., proposed a framework that employs abductive neural networks. They used five simple and weak metrics and by using abductive neural networks selected the best adopted metrics and boosted them. They applied algorithm on PAN 2010 but as it was the case with aforementioned research, their proposed framework suffers from time and memory complexity. Their work also does not take semantic meaning of words into consideration [12]. In another research, El-Alfy deployed lexical similarity metrics and proposed a hybrid method that used three different types of machine learning techniques such as Bayesian learning, support vectors machine and artificial neural networks. The author applied algorithms to MSRPC and demonstrated that artificial neural networks method performed better than Bayesian learning, but it takes a long time to train it [13].

Barrón-Cedeño, et al., adopted an n-gram approach to recognizing suspicious documents. They tested different n size to generate n-gram on METER corpus and finally proposed 2, 3 for n at word level [14]. Kumar and Tripathi using continuous 3-gram that selects the longest substring in a document to detect plagiarism [15]. Zesch, et al., demonstrated that context-based measures cannot detect all forms of text similarity and proposed to apply style-based and grammar-based measures. They came to the conclusion that for more accurate detection, all types of similarity measures should be exploited together. They also stated that text features should be properly selected so that the similarity measure works to its optimum [16]. Both Brockett, et al.[17] and Rus, et al. [18] used combined lexical, semantic and grammatical features in support vector machine to detect paraphrasing.

As it is clear from the review of the literature, few researches have focused on the ambiguity of the word pair's comparison and similarity of structural and style of writing in content similarity measurement; meanwhile, there is a need to use lexical, structural and stylistic features to obtain a comprehensive evaluation of text similarity. To do this, we need to use mathematical theories that have capability to deal vague data. The fuzzy set theory can represent expert knowledge and can deal with ambiguity in real problem. We use this theory to design our new method. In the next section, fundamental concepts of fuzzy sets theory will be discussed.

3. Problem Statement

Text similarity detection can be stated as finding two similar parts of two different documents. Since users change the word order of sentences, style of writing, transpose different sections of a text or rewrite a text by changing a word to its equivalent semantic meaning (substituting hyponyms, synonyms or paraphrasing), detecting these types of text similarity is too difficult, especially when we are concerned with specific content domain texts [7] and in low resource language such as Persian. However in specific content domain texts usually it is difficult to change the style of origin author. This will help us significantly to detect similar document. So we need to clarify the problem statement and identify factors that affect similarities measurement to obtain a better performance. In this paper, we consider the following question:

Given two Persian contents in specific (scientific) content domain, how can we assess the degree of similarity between two texts with high accuracy and precision?

The proposed method uses two kinds of features to solve this problem. It uses lexical features to assessing surface similarity of text pairs and applies structural features to estimating similarity in author writing style and usage of hyponyms words.

4. Fundamental Concepts of Fuzzy Set Theory

Fuzzy sets theory was introduced by Lotfi Zadeh in 1965[19]. This theory is an appropriate framework for handling uncertain and imprecise data. This framework uses a set of "if-then" rules assigned to inference, in which any of these rules are defined by fuzzy sets. Fuzzy logic uses linguistic variables, which can be easily understood by humans and allows decision-making in spite of incomplete and uncertain information.

Fuzzy inference systems can provide appropriate and practical solutions to complex systems engineering in different situations. These systems consist of four main components that seen in figure 2. These components are described as follows [20]:

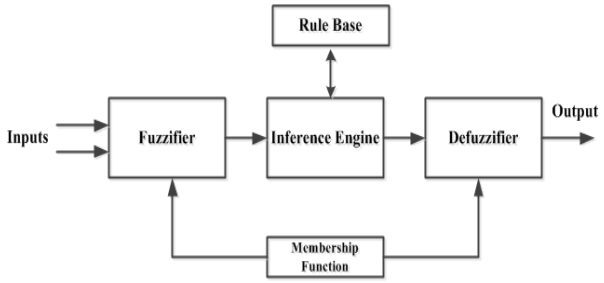


Fig. 1. Basic configuration of fuzzy inference system [20]

In the fuzzifier process, relationships between the inputs of system and linguistic variables are defined by fuzzy membership functions. In this research, each input variables model as trapezoidal fuzzy number that donated as [a,b,c,d] and fuzzy membership are being defined as follows:

$$\mu_A(x)=\begin{cases}\frac{x-a}{b-a} & a\le x\le b\\ 1 & b\le x\le c\\ \frac{x-d}{c-d} & c\le x\le d\\ 0 & o.w.\end{cases} \quad (1)$$

In knowledge base, all linguistic rules are extracted from the domain experts, that is, experts on linguistics. These rules use linguistic variables to express relationships between inputs and outputs of the system. The format of the rule is represented as follows:

If *"the input conditions are true" then "the set of outputs is inference"*.

Inference system is related to the decision part of the system and is able to infer outputs using fuzzy rules and operators. In this research, Mamdani product inference system through the following process generates outputs by using inputs based on predefined rules [21]:

$$\mu_{A^k\to B^k}(x,y)=\mu_{A^k}(x).\mu_{B^k}(y) \quad (2)$$

In which $\mu_{A^k}(x)$ signifies the membership function value of k^{th} rule in the knowledge base.

The defuzzification step performs the reverse of fuzzifier process and generates a crisp value of fuzzy output. There are many techniques to defuzzification. In this research, the center of gravity is exploited in defuzziness process as follows [21]:

$$y'=\frac{\int y_i\mu_B(y)\,dy}{\int \mu_B(y)\,dy} \quad (3)$$

5. The Proposed Method

In this study, a mixed fuzzy inference system (FIS) was proposed which accomplishes the inference through two FISs, assesses similarity between two sentences and finally detects text or document similarity. The advantages of a combination of lexical and structural features are obvious [16]. As in [16] mentioned; experiments show the drawbacks pertaining to use features alone seem complementary and therefore it is good idea to take composing a mixed system combining these two types features. Such combination don't optimize the FIS but it help to improve the accuracy and performance of overall system. Figure 2 demonstrates the conceptual structure of the proposed method. The proposed method is composed of four components: preprocessing, segmentation, features extraction and similarity measures selection and finally fuzzy inference system. These components are described in the rest of the paper.

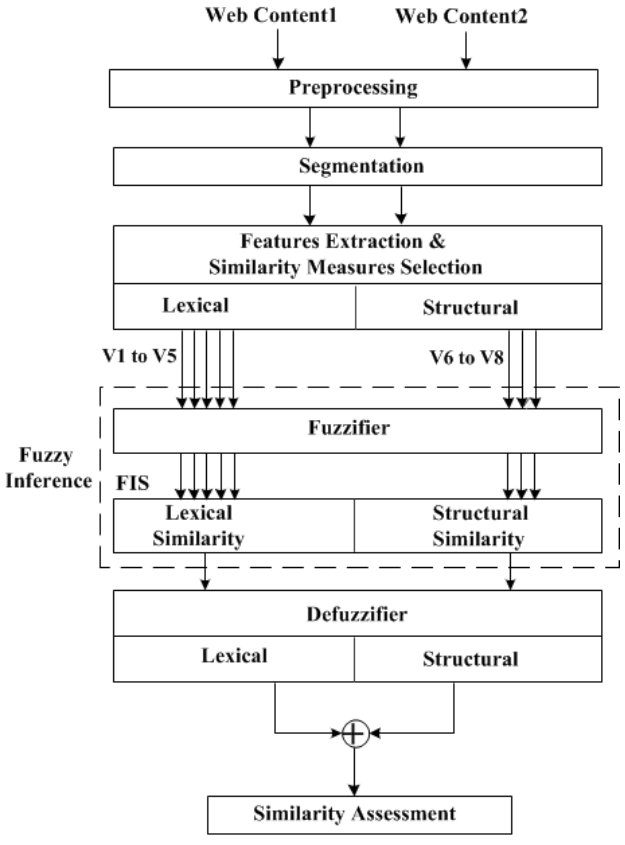


Fig. 2. Proposed mixed FIS Method

5.1 Preprocessing and Segmentation

In these two components, preprocessing operations such as stemming and stop-word removal is done in each input text. Text contains one or more words that express a special meaning in the context in which it is stated. This

meaning is usually expressed through domain-specific knowledge (DSK) terms. Therefore, we need to disambiguate the sense of a word according to its context. To gain more precision and accuracy, in the next step each text is divided into general and domain-specific knowledge parts. In the general part, words such as 'book' and in the domain-specific knowledge part words such as 'E-Learning' are included.

For instance, consider the following two sentences:

1. *"I am working in web programming and semantic web" and*
2. *"Alex has experience in business intelligent and software applications".*

Figure 3 depicts the segmentation of these sentences after preprocessing. Domain-specific knowledge terms and words are extracted using the ontology of IT technical terms. In this research, due to the lack of such domain-specific knowledge ontology in Persian, we had to compile it.

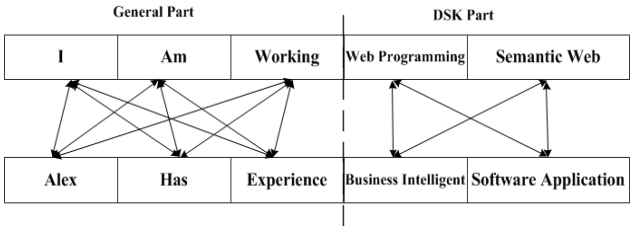


Fig. 3. Segmentation and comparing in sentence pairs

5.2 Measures and Features Extraction

After segmentation, the relevant sections are compared in a way that the general part of the first sentence is being compared to the general part of the second sentence; the same process also occurs for domain-specific knowledge parts. We used lexical- and structural- based features and metrics to assess the extent of similarity between two the texts. For this propose we classified main metrics and measures that mentioned in literature in lexical and structural category. These measures have been experimented on part of PeLeT corpus (20%). To select most appropriate measures, we define *Pr_Ti* criteria as follow:

$$\Pr\_\mathrm{Ti} = \alpha.\,\mathrm{Precision} + (1-\alpha).\,\mathrm{Time\ complexity} \quad (4)$$

Where *Precision* shows the measure accuracy in text similarity detection and *Time complexity* is its time complexity. *α* should be determine based on the problem. In the context of this paper our main goal is accuracy improvement in text similarity detection and therefore we set$\alpha = 1$. Based on the results, *overlapping ratio* and *skip gram* have been selected from lexical category and *stopword overlapping* and *the number of hyponyms words* have been selected from structural category.

The selected features and metrics have been categorized in two groups and are introduced in the following:

5.2.1 Lexical Approach Similarity

The similarity measures and features in this approach only address the surface level of words in contents and do not take the meaning and senses of them into account.

A. Overlapping Ratio.

This measure calculates the common words between two texts by using n-grams in word or character level, divided by the length of the first and second text respectively. Then, it computes the geometric mean of the two ratios. This ratio for first text computes as equation (5) [22].

$$S(T_1,T_2)=\frac{|T_1\cap T_2|}{|T_1|} \quad (5)$$

Which $|T_1 \cap T_2|$ is a number of common words between two texts and $|T_1|$ is a number of first text words.

This measure is to be handled in general and DSK parts separately and for simplicity denoted by V1, V3 respectively in this paper.

B. Skip-Gram Measure

This measure is similar to the n-gram but can skip between words as 'skip' length. According to the examination on Persian IT corpus, the best skip number was found to be 3. This measure deals with word order as well as detects the common phrases in two sentences [23].This measure also uses general and DSK parts separately and is denoted by V2, V4 respectively.

C. The Ratio of Number of DSK to General Words in Union of Two Texts

This ratio indicates the degree of importance of DSK words to general words in union of two texts and is denoted by V5.

5.2.2 Structural Approach to Similarity

In this approach that can be considered the same as the lexical approach, the structural features of each text are considered as follows:

A. The Number of Hyponyms Words

This feature has been selected for the reason that in plagiarized texts the hyponyms words such as "look and view" are often used interchangeably for secrecy. This feature is handled in general and DSK parts separately and denoted by V6, V7 respectively.

B. The Overlapping of Stopword

This feature has been chosen because if two texts are structurally similar, the same structure of word stop is being used specially in scientific texts. This feature can be considered as author style that is denoted by V8[16].

Table 1 presents the complete list of measures and features used in fuzzy systems. The linguistic variables of the proposed FIS were developed on the basis of these measures and features.

Table 1. Selected Features and Measures

		Union of Two Texts	Term Part	
			General	Domain Specific Knowledge
Similarity Approach	Lexical	The ratio of DSK words to general words	Overlapping Ratio Skip Gram	Overlapping Ratio Skip Gram
	Structural	The Overlapping of Stopword	The Number of Hyponyms Words	The Number of Hyponyms Words

5.3 The Surface level Text Similarity Fuzzy Inference System

Due to the complex nature of natural language, it is quite difficult to detect similarity between scientific and domain-specific knowledge texts. Meanwhile, the ambiguity inherent in human language prohibits us from developing efficient NLP techniques. By the same token, the same content might be worded differently in various paraphrases. That is why we need to gather information from experts to achieve a more precise assessment. To do this, we designed two lexical and structural FISs that assess text similarity from a different perspective. This is our first contribution. The output of each FIS will be combined and finally the proposed mixed method can determine whether two sentences like each other base on weighted average of two lexical and structural FISs result. We model similarity assessment as fuzzy linguistic variables to overcome the ambiguity and vagueness of the assessment. In Table 2 similarity of linguistic variables and their membership function will be expressed.

Table 2. The Similarity of Linguistic Variable and Membership Function

Linguistic Variable	Interval
Different	(-∞,0,0.1,0.35)
Semi Similar	(0.1,0.35,0.55,0.75)
Similar	(0.55,0.75, 1,+∞)

Since assessing the similarities between the two texts is defined as a fuzzy number, inputs of FIS are also modeled as a fuzzy number. For easy and quick access, Table 3 displays all types of variables and Table 4 depicts variables used in each FIS with their correspondence.

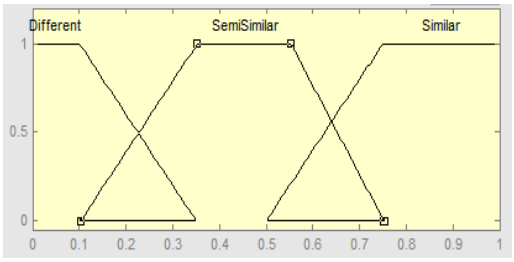
Table 3. Lingustic Variables in Simialrity Assesment

TYPE	LINGUISTIC VARIABLE	INTERVAL
T1 (Stopword overlaping)	Low	(-∞,0,0.1,0.35)
	Middle	(0.1,0.35,0.55,0.85)
	High	(0.6,0.75, 1,+∞)
T2 (Hyponymword overlaping)	Low	(-∞,0,0.1,0.3)
	Middle	(0.1,0.3,0.5,0.7)
	High	(0.6,0.7, 1,+∞)
T3 (The ratio of DSK words to general words)	Low	(-∞,0,0.1,0.25)
	Middle	(0.15,0.35,0.55,0.8)
	High	(0. 5,0.75, 1,+∞)
T4 (Similarity Assessment)	Different	(-∞,0,0.1,0.35)
	Semi Similar	(0.1,0.35,0.55,0.75)
	Similar	(0.55,0.75, 1,+∞)

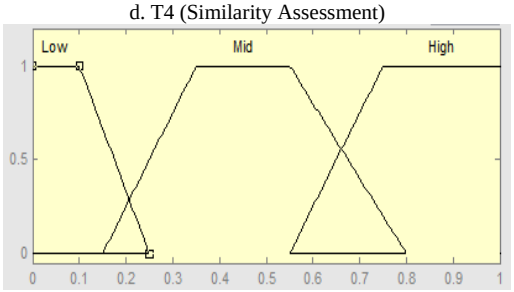
Table 4. FISs Variables

FIS	PART	VARIABLE NAME	VARIABLE TYPE
Lexical Similarity	General	V1	T4
		V2	T4
	DSK	V3	T4
		V4	T4
	Union of two Texts	V5	T3
Structural Similarity	General	V6	T2
	DSK	V7	T2
	Union of two Texts	V8	T1

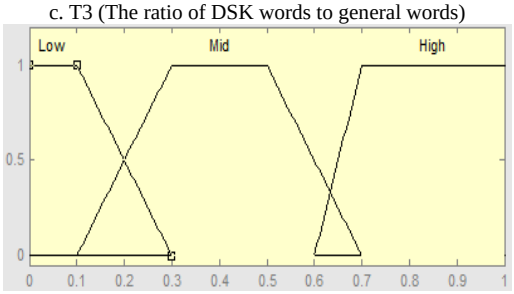
Because of wide variety of ambiguity in word or text pairs comparisons; all fuzzy variables model as trapezoidal membership function. Figure 4 also displays the membership function of variable types in FISs.



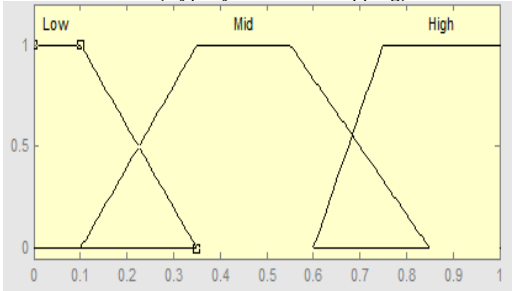
d. T4 (Similarity Assessment)



c. T3 (The ratio of DSK words to general words)



b. T1 (Hyponymword overlapping)



a. T1 (Stopword overlapping)

Fig. 4. The membership functions of FISs variables

We implemented mixed similarity fuzzy inference system using MATLAB 2010 fuzzy tool and used Mamdani type of fuzzy system with following configuration that set by try and error:

- AND method: prod
- OR method: max
- Defuzzification method:LOM (Large Of Maximum)

5.4 Fuzzy Function of Part Pairs Similarity

Since each part is comprised of several words, we need to compare word pairs and then aggregate their results as it was done for the example in Figure 2. The only difference in this section is that the comparisons will be made in a fuzzy manner. As a result, for example, the output matrix resembles equation 4. Each element of this matrix is a fuzzy number, in which we use maximum as S norm in each row and column to compute the final result of two texts; then, the values of the average of these two fuzzy numbers are obtained by Rosenfeld relationship [21].

$$\mathrm{Sim}_{\mathrm{DSK}}(\mathrm{T}_1, \mathrm{T}_2) = \begin{bmatrix} \{(\mathrm{Similar}, 0.53), (\mathrm{Semi-similar}, 0.12), (\mathrm{Dissimilar}, 0.16)\} & \{(\mathrm{Similar}, 0.73), (\mathrm{Semi-similar}, 0.4), (\mathrm{Dissimilar}, 0.08)\} \\ \{(\mathrm{Similar}, 0.81), (\mathrm{Semi-similar}, 0.32), (\mathrm{Dissimilar}, 0.11)\} & \{(\mathrm{Similar}, 0.12), (\mathrm{Semi-similar}, 0.22), (\mathrm{Dissimilar}, 0.67)\} \end{bmatrix} \quad (6)$$

5.5 Fuzzy Rules Base

After precisely defining FIS variables, system rules were deduced and developed by conducting an interview with a group of five natural language and domain experts. Table 5 and 6 demonstrates some fuzzy rules for lexical and semantic FIS respectively. In this rule base as mentioned, we consider fuzzy "AND" for t-norm and operators among fuzzy variables and the operator for rules aggregation is s-norm.

For example, rule 5 in Table 5 is as follow:

If *"Overlapping Ratio [V2] in General Part is Low"*, AND *"Overlapping Ratio [V4] in Knowledge domain Part is Low"*, THEN *"[Lexical] Similarity is Different"*.

Table 5. Some rules of FISs

FIS	RULE #	VARIABLE NAME					SIMILARITY
		V1	**V2**	**V3**	**V4**	**V5**	
Lexical similarity	1	-	High	-	Low	Low	Similar
	2	-	High	Low	Low	Middle	Semi Similar
	3	-	High	Middle	Low	Middle	Semi Similar
	4	Low	Middle	Low	Low	Middle	Different
	5	-	Low	-	Low	-	Different
Structural similarity		**V6**	**V7**	**V8**			
	1	Low	Low	Low			Different
	2	Middle	Low	Low			Different
	3	-	High	-			Similar
	4	Middle	Low	Middle			Semi Similar
	5	High	Low	Middle			Semi Similar

6. Numerical Results

In this section, the numerical results of the proposed method will be reported. The proposed method was implemented in Visual Studio Environment and used MATLAB fuzzy toolbox to simulate FISs. Because of there is no technical English corpus, we applied the method on Persian corpus to evaluate its efficacy. To create the Persian corpus, that named PeLeT, we used the papers' abstracts presented in ICELET conferences in e-learning domain knowledge. This corpus contains 810 sentence pairs totally that each *"Different"*, *"Semi Similar"* and *"Similar"* class has 270 sentence pairs. The first sentence of pairs gathered from papers' abstracts presented in ICELET and a group of expert domain generated a paired sentence in one of the randomly assigned class. The average length of sentence is 13 words. After preprocessing and segmentation, the proposed method applied to the corpus. Table 6 indicates some examples of text pairs in PeLeT corpus.

Table 6. Some examples of text pairs in PeLeT corpus

FIRST TEXT	PAIRED TEXT	CLASS
E-learning is a new type of learning using information technology tools.	The expansion of changes in information technology led to a new type of learning called e-learning.	Similar
The most important goal of e-learning is to transfer the focus of learning from teacher to learner.	One of the most important goals of e-learning is to create a learner-centered atmosphere.	Semi-similar
The intelligent educational system adapts the educational strategy to the learner characteristics.	The main part of the intelligent educational system is the learner-based model that includes information which the system holds about the learner. The educational strategy is adapted based on the information obtained from this section.	Different

In final step, two FISs outputs fused and this fusion is our other contribution. For lexical and structural similarity detection FISs, weights 0.65 and 0.35 have been considered respectively. Table 7 shows results of two FISs fusion for two sample text in corpus.

Table 7. Fusion of lexical and structural similarity detection FISs

FIS	SIMILARITY ASSESSMENT			WEIGHTS
	Similar	Semi-Similar	Different	
LEXICAL	0.73	0.16	0.09	0.65
STRUCTURAL	0.57	0.25	0.12	0.35
FUSION	0.674	0.192	0.205	

Table 8 indicates the confusion matrix of the proposed method. The first row of table indicates that our proposed method correctly detect 185 of 270 sentence pairs are similar. This method also failed for rest of sentence pair and placed 20 and 56 of sentence pairs in semi-similar and different class respectively. The other rows show the same results.

Table 8. Confusion matrix of proposed method

		Predicted Class		
		Similar	Semi Similar	Different
Actual Class	Similar	185	20	56
	Semi Similar	28	173	78
	Different	41	53	176

To evaluate the performance of the proposed method, we used recall, precision and F-measure that are widely employed in text mining. It should be noted the use of other languages datasets such as WordNet will not show the method efficiency.so we also intended to cast more light on the comparison of the efficiency of the proposed algorithm and given the fact that the proposed method has been applied to PeLeT Corpus. We experimented three methods mentioned in the literature with PeLeT Corpus to assess the efficiency of our method. Table 9 illustrates the results of proposed method implementation and its comparison with three methods. The Cosine Coefficient was considered as baseline. The results show that proposed method outperforms the other methods.

Table 9. The result performance

Method	F Measure	Percision	Recall
The Proposed Method	0.78	0.75	0.81
Gupta, et al.[11]	0.75	0.72	0.77
Alzahrani and Salimi[10]	0.65	0.63	0.67
Mihalcea and Corley[26]	0.62	0.60	0.65
Cosine Coefficient (Baseline)	0.55	0.54	0.57

7. Conclusions

In this paper, a mixed fuzzy inference system method was proposed to overcome the ambiguity and consider structural features and author style of writing in the similarity measurement in Persian texts. In this method, in the first step, after preprocessing and stop word removal, the text is divided into general and domain-specific knowledge parts then appropriate features are extracted and similarity metrics are calculated. Two lexical and structural fuzzy inference systems were designed that its rules are extracted from the experts' knowledge and finally the outputs of these two FISs are integrated through weighted combination. With regard to the fact that the proposed method was applied to PeLeT Corpus, we also carried out tests using three methods proposed in the literature to evaluate its efficiency. Such a comparison was thought to throw light on the efficiency of the proposed algorithm. The results show that the proposed method outperforms than others and gained accuracy rate of 78% which increases precision and recall measure.

References

[1] Wang, Yong, and Julia Hodges. "Document clustering with semantic analysis." System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on. Vol. 3. IEEE, 2006.

[2] Mohler. M., , and Mihalcea Rada. "Text-to-text semantic similarity for automatic short answer grading." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[3] Hariharan. S., , et al. "Detecting plagiarism in text documents." Information Processing and Management. Springer Berlin Heidelberg, 2010. 497-500.

[4] Barrón-Cedeño, Alberto, and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison." In Advances in Information Retrieval, pp. 696-700. Springer Berlin Heidelberg, 2009.

[5] Kent, Chow Kok, and Naomie Salim. "Features based text similarity detection." arXiv preprint arXiv:1001.3487 (2010).

[6] Joy, Mike, and Michael Luck. "Plagiarism in programming assignments." Education, IEEE Transactions on 42.2 (1999): 129-133.

[7] Potthast, Martin, et al. "Cross-language plagiarism detection." Language Resources and Evaluation 45.1 (2011): 45-62.

[8] Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42.2 (2012): 133-149.

[9] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12, no. 5 (2012): 1493-1502.

[10] Alzahrani, Salha, and Naomie Salim. "Fuzzy semantic-based string similarity for extrinsic plagiarism detection." Braschler and Harman (2010).

[11] Gupta, Rohit, et al. "UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment." SemEval 2014 (2014): 785.

[12] El-Alfy, El-Sayed M., et al. "Boosting paraphrase detection through textual similarity metrics with abductive networks." *Applied Soft Computing* 26 (2015): 444-453

[13] El-Alfy, El-Sayed M. "Statistical Analysis of ML-Based Paraphrase Detectors with Lexical Similarity Metrics." In *Information Science and Applications (ICISA), 2014 International Conference on*, pp. 1-5. IEEE, 2014.

[14] Barrón-Cedeño, Alberto, and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison." In Advances in Information Retrieval, pp. 696-700. Springer Berlin Heidelberg, 2009.
[15] Kumar, Ranjeet, and R. C. Tripathi. "A Trigram Word Selection Methodology to Detect Textual Similarity with Comparative Analysis of Similar Techniques." In *Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on*, pp. 383-387. IEEE, 2014.
[16] Zesch, Daniel Bär1 Torsten, and Iryna Gurevych. "Text reuse detection using a composition of text similarity measures." In *Proceedings of COLING*, vol. 1, pp. 167-184. 2012.
[17] Brockett, Chris, and William B. Dolan. "Support vector machines for paraphrase identification and corpus construction." In *Proceedings of the 3rd International Workshop on Paraphrasing*, pp. 1-8. 2005.
[18] Rus, Vasile, Philip M. McCarthy, Mihai C. Lintean, Danielle S. McNamara, and Arthur C. Graesser. "Paraphrase Identification with Lexico-Syntactic Graph Subsumption." In *FLAIRS conference*, pp. 201-206. 2008.
[19] Zadeh, Lotfi A. The concept of a linguistic variable and its application to approximate reasoning. Springer US, 1974.
[20] Huang, Yo-Ping, et al. "An intelligent approach to detecting the bad credit card accounts." Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications. ACTA Press, 2007.
[21] Rutkowski, Leszek, and Krzysztof Cpalka. "Flexible neuro-fuzzy systems." *Neural Networks, IEEE Transactions on* 14.3 (2003): 554-574.
[22] Metzler, Donald, et al. "Similarity measures for tracking information flow." Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.
[23] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[24] Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994.
[25] Resnik, Philip. "Using information content to evaluate semantic similarity in a taxonomy." *arXiv preprint cmp-lg/9511007* (1995).
[26] Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." *AAAI*. Vol. 6. 2006.

Hamid Ahangarbahan is currently a Ph.D. candidate in Information Technology Engineering in Tarbiat Modares University, Tehran, Iran. His research interests are in Soft computing, computational intelligence, Information Engineering and Data Mining Knowledge Discovery, Intelligent Methods and System Modeling.

Gholam Ali Montazer received his B.Sc. degree in Electrical Engineering from Kh. N. Toosi University of Technology, Tehran, Iran, in 1991, his M.Sc. degree in Electrical Engineering from Tarbiat Modares University, Tehran, Iran, in 1994, and his Ph.D. degree in Electrical Engineering from the same university, in 1998. He is a Professor of the Department of Information Technology Engineering in Tarbiat Modares University, Tehran, Iran. His areas of research include Information Engineering, Knowledge Discovery, Intelligent Methods, System Modeling, E-Learning and Image Mining.