Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm
2015 Vol.5 (S1), pp. 3576-3581/Shahabi
Research Article

USING FUZZY CLUSTERING TO DETECT SEMANTIC SIMILARITY AMONG PERSIAN SENTENCES

***Amir Shahab Shahabi**
Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran, P.O. Box 11365/4435
**Author for Correspondence*

ABSTRACT
Due to the fact that in the final summary, the repetition of the same sentences is to be eliminated, distinguishing the existing similarities both sentences and paragraphs may be highly required in the context of Multi-Document summarization. Therefore, to help applying such anti-redundancy, we need a mechanism which effectively determines the possible semantic similarities between each pair of sentences and/or expressions and eventually between each pair of texts arises. Thus, this paper utilizes a fuzzy approach aiming the semantic similarities. Hence, to pinpoint the possible similarities, the article makes use of both fuzzy similarity and fuzzy approximation relations. The proposed approach begins with obtaining the lemma of Persian words and verbs and a fuzzy similarity relation generated afterwards by the synonyms. Using both generated fuzzy similarity relation and fuzzy proximity relation, the sentences with nearly same meanings are detected. The results indicated notable redundancy-free final summary which has kept valuable information presented in original documents.

Keywords: *Multi-Document Summarizer, Fuzzy Similarity Relation, Fuzzy Proximity Relation, Lemma, Fuzzy Relations Composition, Anti-Redundancy, Syntax Parser, Meta Variable, Meta Rule, Paradigmatic, Tokenizer, Lemmatizer*

INTRODUCTION
Despite single document summarization, Multi-Document summarization requires summarizers be able toprecisely distinguish the similar sentences and texts in order to achieve the anti-redundancy factor which, by the way, plays the most significant role in Multi-Document Summarization (Goldstein *et al.*, 2000). Therefore, various approaches have been proposed in the literature. Some fuzzy based approaches including (NatelKhanlari, 1991) and (Aboumahboob, 1996) tried to distinguish the similarity of two sentences in Persian language via their concept, meaning of words, expressions, noun phrases, and verb phrases utilized in those sentences.

Related Works
In this paper, for effectively finding the similarity, grammar, tokenize and parser proposed by Shahabi (1997) is utilized to tell words, nouns, and verb phrases apart from each other in a Persian text. Afterwards, using lemmatizing proposed by (NatelKhanlari, 1991), (Siemens, 1996), (Dichy *et al.*, 2001), (Bateni, 1992), the lemma of words and verbs is found. According to the fact that meanings of the words are completely context dependent, special knowledge base is required to accurately obtain the meanings of the words. Since a fuzzy similarity relation is created by all the words that can be substituted with their synonyms on a paradigmatic relation basis (Zimmermann, 1996), (Wang, 1997), the required knowledge base is generated by the aforementioned relation. Furthermore, fuzzy relations generated for each sentence in the text help the system to determine the possible similarities betweeneach pair of sentences. Using each relation derived from each sentence accompanied by the knowledge base helps to conclude a brand-new relation that clearly indicates which words are knowledge based and accordingly which words can be substituted by their synonyms. On the other hand, since fuzzy relations are obtained for all sentences in the text, a fuzzy proximity relation is generated by selecting a pair of obtained fuzzy relations. Fortunately, fuzzy proximity relation makes determining the similarity between each pair of sentences possible (Dubois and Prade, 1980), (Fujimato and Sugano, 1997). Repeatedly generating fuzzy proximity relations for each pair of previously generated fuzzy relations results in clustering the sentences

based on their meanings. At the end, this seems appropriate to mention that clustering the sentences is performed using $\alpha - cut$ rule (Marcu and Gerber, 2001). Some recent work extends multi-document summarization to multilingual environments (Evans, 2005). In another work the idea of multilayered text similarity graph is presented based on it the authors present a modified iterative graph-based sentence ranking algorithm (Canhasi, 2014). Although much work has been done to eliminate the redundancy in multi-document summarization, the problem is still actual and addressed in the work as well. The work proposes to integrate the generalized BFOS algorithm adopted for pruned tree structured quantize design with the HAC (Hierarchical Agglomerative Clustering) algorithm. The two main parameters (distortion and rate) in the latter work are adoptedfrom the multi-document summarization task. Distortion can be succinctly defined as the information loss in the meaning of the sentences due to their representation with other sentences. (Attokurov and Bayazit, 2014).

Text Tokenizing and Syntax Parsing

To extract the correct meanings of the words in a text from corpus, the first step is to figure out their correct part of speech i.e. noun, verb, noun phrase, or verb phrase. To do so, a Persian language grammar based tokenizer and syntactic parser is utilized. Hence, a suitable grammar is absolutely required. As it is generally known that the grammar of a natural language is quite unrestricted, wrongfully parsing a sentence due to ambiguity and consequently obtaining several parse trees for a specific sentence are expectedly reasonable. To overcome such problems, a *two level grammar* is generated which is an unambiguous context free version of natural language grammar. The *two level grammars* involve some Meta variables whose initialization produces a context free grammar that is considerably much easier to parse (Krulee, 1991). This should be mentioned that a set of rules are required to determine how to initialize aforementioned Meta variables of the *two level grammar*. The lack of such rules makes covering a wide area of a language hardly possible.

Lemmatizing Persian Words

Lemmatization is the function of reducing the overhead of the words and extracting their respective roots or lemma. As of the root of a word is accurately acquired, the meaning of that word becomes sufficiently convenient (Siemens, 1996). According to (Dichy *et al.*, 2001), Persian and Arabic words might have four different overhead types including:

1. Enclitics – objective connected pronouns like BICHAREAM that the lemma is BICHARE (means poor) (NatelKhanlari, 1991).
2. Suffixes – plural sign or relative adjective signs like BARG HA that BARG is the lemma of it or IRANI that its lemma is IRAN.
3. Proclitic – like AL in Arabic words.
4. Prefixes – that can be noun, adjective or pronouns like HAMANDISHI that its lemma is ANDISHE.

Knowledge Base Creation for Synonym Words

As it was mentioned before, the knowledge base of synonym words is a fuzzy relation. Indicating Was our universal set of all the words in a text, and knowing that each word in W takes one of the forms of noun, adjective, verb or any phrasal expression, the aim is to find synonyms(i.e. those words that can be substituted with each other) of the words in the sentences (Aboumahboob, 1996). To do so, a fuzzy relation between W and itself is needed (Zimmermann, 1996). As shown below, $\tilde{P}$ (the first letter of the word *Paradigmatic*) represents the fuzzy relation between W and itself.

$$\tilde{P} = \{((w_1, w_2), \mu_{\tilde{P}}(w_1, w_2)) \mid (w_1, w_2) \in W \times W\}$$

Where w_1 and w_2 indicate the words in Persian language. The membership function ($\mu_{\tilde{P}}$) of $\tilde{P}$ is shown as:

$$\mu_{\tilde{P}}(w_1, w_2)$$

Where $0 < \mu_{\tilde{P}} < 1$ and $\mu_{\tilde{P}}$ depends on how much w_1 and w_2 are near each other.

The rest of the paper is clearly explained using the following example.

Example1. Assume that there are three different sentences as shown below:

– S1. Students go to school at educational year.
– S2. Students present in class at fall.
– S3. Lessons stated by instructors should have been learned by students.

Since all the sentences have a similarity relation in meaning, the aim is to find that similarity. Nonetheless, all the words used in the sentences above are related to each other via a membership function whose value, which should be determined by a literature specialist, indicates the semantic similarity among them.

Considering aforementioned assumption, the knowledge base and synonyms of the utilized words are to be defined. Therefore, the word and phrase set for the aforementioned sentences is:

$W = \{student, togo, school, educationa lyear, topresent, class, fall, lesson, tostate, instructor, tolearn\}$

And the fuzzy relation that specifies the knowledge base is shown in Table 1:

Table 1: Fuzzy Relation $\tilde{P}$ for W

	Student	To go	school	Educational year	To present	class	Fall	lesson	To state	instructor	To Learn
Student	1	0	0	0	0	0	0	0	0	0	0
To go	0	1	0	0	0.7	0	0	0	0	0	0
School	0	0	1	0	0	0.8	0	0	0	0	0
Educational Year	0	0	0	1	0	0	0.9	0	0	0	0
To present	0	0.7	0	0	1	0	0	0	0	0	0
Class	0	0	0.8	0	0	1	0	0	0	0	0
Fall	0	0	0	0.9	0	0	1	0	0	0	0
Lesson	0	0	0	0	0	0	0	1	0	0	0
To state	0	0	0	0	0	0	0	0	1	0	0
Instructor	0	0	0	0	0	0	0	0	0	1	0
To learn	0	0	0	0	0	0	0	0	0	0	1

Distinguishing Similarity Relation among Sentences

This begins by generating fuzzy relation for each of the sentences. Each fuzzy relation is a vector with n components where $n = |W|$. That is, each fuzzy relation represents each corresponding sentence with all the words in the knowledge base. The value of the membership function for each existing word in the sentence equals 1 and for those non-existing words is 0. Table 2 shows fuzzy relations for each of the aforementioned sentences.

Table 2: Fuzzy Relation of sentences S1, S2, S3, and S4

		Student	To go	School	Educational year	To present	class	fall	lesson	To State	instructor	To Learn
$\tilde{R}_1$	S1	1	1	1	1	0	0	0	0	0	0	0
$\tilde{R}_2$	S2	1	0	0	0	1	1	1	0	0	0	0
$\tilde{R}_3$	S3	1	0	0	0	0	0	0	1	1	1	1

After generating fuzzy relations, it should be determined which words in knowledge base have substitutions in the sentence. To do so, combining each ongoing sentence's fuzzy relations with the relations representing the knowledge base seems possible. This should be indicated that the membership

value of the substitutable words in the sentence is between 0 and 1. Therefore, in the example, the aforementioned composition of fuzzy relations is a fuzzy min-max composition of relations $\tilde{R}_1$, $\tilde{R}_2$ and $\tilde{R}_3$ with $\tilde{P}$. Table 3 shows the results of the combining fuzzy relations with knowledge base relation.

Table 3: Fuzzy Max-Min Composition of sentences' fuzzy relations with knowledge base relation

		Student	To go	School	Educational year	To present	Class	Fall	Lesson	To state	Instructor	To Learn
$\tilde{R}_1 \circ \tilde{P}$	S1	1	1	1	1	0.7	0.8	0.9	0	0	0	0
$\tilde{R}_2 \circ \tilde{P}$	S2	1	0.7	0.8	0.9	1	1	1	0	0	0	0
$\tilde{R}_3 \circ \tilde{P}$	S3	1	0	0	0	0	0	0	1	1	1	1

To precisely detect the similarity that the sentences share, fuzzy proximity relation is used taking account of the previously generated fuzzy relations for each sentence. The name of this relation is fuzzy tolerance relation (Dubois and Prade, 1980). While this relation is supposed to be flexible and symmetric, this relation becomes a similarity relation if transitive property is added to it. Therefore, the relation for two sets of $X = \{x_1, x_2, ...\}, Y = \{y_1, y_2, ...\}$ is defined as follows (Fujimato and Sugano, 1997):

$$S = \frac{| R_{y_i} \cap R_{y_j} |}{\min\{| R_{y_i} |, | R_{y_j} |\}}$$

Where R_{y_i} isa subset of X that relates with y_i and R_{y_j} is a set or subset of Y s that relates with y_j. S isthe similarity between R_{y_i} and R_{y_j}.

This should be mentioned that while $\tilde{A}$ is a fuzzy set, then $|\tilde{A}|$ is the cardinality of $\tilde{A}$ which is calculated as follows (Wang, 1997), (Zimmermann, 1996):

$$|\tilde{A}| = \sum_{i=1}^{n} \mu_{\tilde{A}}(x_i)$$

Accordingly, S is the cardinality of intersection of R_{y_i} and R_{y_j} divide by minimum of cardinality of R_{y_i} and R_{y_j}. Since S relation is reflexive and symmetric, not only is it considered a fuzzy proximity relation, but also the relation to distinguish the similarity between each two sentences. In case of the aforementioned sentences, the fuzzy proximity relations between each pair of them are:

$$S_{12} = \frac{5.8}{6.4} = 0.90625$$

$$S_{13} = \frac{1}{5} = 0.2$$

$$S_{23} = \frac{1}{5} = 0.2$$

As it is obvious, the S_{12} is greater than the other two relation values. So,it is construed that the similarity between first and second sentences is more than any other pair of sentences.

As it was previously stated that clustering the sentences is performed using $\alpha - cut$. Therefore, clustering is done via a fuzzy similarity relation like $S \geq S_\alpha$ based on a suitable $\alpha - cut$ that has shown a considerable improvement in multi-document summarizing system.

RESULTS AND DISCUSSION

Results

The proposed mechanism is tested through a text with 58 sentences. According to a human specialist, the testing text contains 15 clusters of the sentences with approximately the same meanings. All the information about each cluster including number of the sentences in each cluster and their normal weights are shown in Table 4. Applying the proposed knowledge-based mechanism containing 946 words and synonyms and considering suitable $\alpha - cut$ as $S_\alpha = 0.7$, 22 clusters are found.

Table 4: Results of performing system run on a text with 58 sentences

Text clusters Based on Human specialist Detection	Number of Sentences Per Cluster	Normal Weight Of a Cluster	Number of Sentences per Cluster made By system	Error rate Per Cluster
C1	9	0.9*1/15	7	22.2%
C2	6	0.6*1/15	6	0%
C3	10	1.0*1/15	5	50%
C4	4	0.4*1/15	4	0%
C5	3	0.3*1/15	2	33.3%
C6	8	0.8*1/15	8	0%
C7	9	0.9*1/15	7	22.2%
C8	1	0.1*1/15	2	50%
C9	1	0.1*1/15	1	0%
C10	1	0.1*1/15	1	0%
C11	2	0.2*1/15	2	0%
C12	1	0.1*1/15	2	50%
C13	1	0.1*1/15	2	50%
C14	1	0.1*1/15	1	0%
C15	1	0.1*1/15	1	0%

The weighted average error rate of the proposed mechanism is calculated based on the cluster weights. Which in this case is 1/15*[22.2*0.9+50*1+33.3*0.3+22.2*0.9+50*0.1+50*0.1+50*0.1] = 7.66. In the other words, the proposed mechanism works with 92.34% accuracy.

Discussion

This article attempted to use fuzzy proximity relation to segment the text. Moreover, according to the results, it is concluded that the more α value in S_α increases and the nearer to 1, the more system error decreases. However, the proposed approach is tested using $S_\alpha = 0.7$ due to the fact that generating knowledge base is performed in the company of inevitable errors in determining fuzzy memberships between each pair of words and phrases which per se increase the overall average error rate. Thus, it is tried to eliminate the impacts of those inevitable errors on the results by setting S_α to 0.7 based on empirical considerations.

Conclusion

The summarization mechanism proposed in this paper attempts to solve the problem of finding sentences with nearly same meaning based on a paradigmatic relation. In other words, substituting words in a specific text with their synonyms helps precisely to detect similarity between each pair of sentencesas well as efficiently avoiding redundancy by eliminating approximately equal meaning sentences but one to place in summarization result.

ACKNOWLEDGMENT
The author is pleased to thank Islamic Azad University – South Tehran Branch for its financial support based on grant number b/16/709-91/4/20.

REFERENCES
Aboumahboob A (1996). *Farsi Language Structure* (Mitra Pub.).
Attokurov U and Bayazit U (2014). Multi-document summarization using distortion-rate ratio. *Proceeding of the ACL 2014 Student Research Workshop, Baltimore, Maryland, USA* 64-70.
Bateni MR (1992). *Language Grammar a New Look* (Agah Pub.).
Canhasi E (2014). Multi-layered graph-based multi-document summarization model.
Dichy J, Krauwer S and Yaseen M (2001). On Lemmatization in Arabic, A formal Definition of Arabic Entries of Multilingual Lexical Databases. *Proceedings of the workshop on Arabic language Processing: Status and Prospects 20-30, July 6th, 2001, Association for Computational Linguistics 39th Annual Meeting and 10th Conference of European Chapter, Toulouse.*
Dubois D and Prade H (1980). *Fuzzy Sets and Systems Theory and Applications* (Academic press Inc.)
Evans DK (2005). Similarity-based multilingual multi-document summarization. *Technical Report CUCS-014-05*, Columbia University.
Fujimato T and Sugano M (1997). *Clustering Verb, Adjective, Adjectival Verb Concepts using Proximity Relation* (IEEE).
Goldstein J, Mittal V, Carbonell J and Callan J (2000). Creating and Evaluating Multi-Document Sentence Extract Summaries, *Proceedings of the 2000 CIKM International Conference of Information and Knowledge Management. Mclean VA, USA* 165-172.
Krulee GK (1991). *Computer Processing of Natural Language* (Printice Hall Inc.)
Marcu D and Gerber L (2001). An Inquiry in to the Nature of Multi-Document Abstract, Extracts and their Evaluation. *Proceedings of Automatic Summarization Workshop.*
NatelKhanlari P (1991). *Farsi Language Grammar* (Toos Pub.).
Shahabi A Sh (1997). Farsi Text Understanding. *MS Dissertation.*
Siemens RG (1996). Lemmatization and Parsing with TACT Preprocessing Programs. Department of English University of British Columbia.
Wang LX (1997). *A Course on Fuzzy Systems and Control* (Printice Hall Inc.).
Zimmermann HJ (1996). *Fuzzy Set Theory and its Application*, third edition (Kluwer Academic Pub.).