Information Science and Technology
ISSN: 0976-917X & ISSN: 0976-9188, Volume 3, Issue 1, 2014, pp.-036-038.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BIJ0000002

SIMILAR PERSIAN SENTENCES FUZZY CLUSTERING

SHAHABI A.S.*
Department of Computer Software Engineering, South Tehran Branch-Islamic Azad University, Tehran, Iran.
*Corresponding Author: Email- shahabi_amir@azad.ac.ir

Received: May 17, 2012; Accepted: March 06, 2014

Abstract- Multi-Document summarization strictly needs distinguishing the similarity between sentences & paragraphs of texts because repeated sentences shouldn't exist in final summary so in order to applying this anti-redundancy we need a mechanism that can determining semantic similarities between sentences and expressions and paragraphs and finally between texts. In this paper it's used a fuzzy approach to determining this semantic similarity. We use fuzzy similarity and fuzzy proximity relation for gaining this goal. At first, lemma of Persian words and verbs obtained and then synonyms create a fuzzy similarity relation and via that relation the sentences with near meaning calculated with help of fuzzy proximity relation. So we can produce an anti-redundant final summary that have more valuable information.

Keywords- Fuzzy Similarity Relation, Fuzzy Proximity Relation, Lemma, Anti-Redundancy, Tokenizer, Lemmatizer

Citation: Shahabi A.S. (2014) Similar Persian Sentences Fuzzy Clustering, Information Science and Technology, ISSN: 0976-917X & ISSN: 0976-9188, Volume 3, Issue 1, pp.-036-038.

Introduction

In a Multi-Document Summarizer opposite of a single document summarizer there exist a great need to distinguish of similar sentences & texts in order to achieving the anti-redundancy factor that is one of the most important factors in Multi-Document Summarization [1-4]. For obtaining this goal many different efforts has been done that one of them is discussed in this paper. At this discussion a fuzzy approach used in order to distinguishing similarity of two sentences via their concept. This effort is done for Persian language and is based on concept and meaning of words, expressions, noun phrases and verb phrases in Persian language [5,6]. For this job we should distinguish word and noun and verb phrases from a Persian text that is done by a grammar, tokenizer and parser [7]. After finding words and nouns and verb phrases by tokenizer and syntactic parser the lemma of words and verbs is created by lemmatizer [5,7-12]. Then for determining the meaning of the words we need to a special knowledge base. This knowledge base is created by a fuzzy relation. All words that can be substituted with their synonyms based on a paradigmatic relation, create a fuzzy similarity relation [13,14] and this relation creates our knowledge base. Then creating a fuzzy relation for any sentence in the text makes system capable of determining similarity between sentences via fuzzy relations composition. With compositing a relation of a sentence by our knowledge base we can conclude a new relation that tell us in a sentence which words from knowledgebase exist and which words can be substituted with their synonyms. We do this job for all sentences in the text and obtain a fuzzy relation for each sentence then select a pair of these relations and create a fuzzy proximity relation for them and then we can determine the similarity between those [15,16]. Repeating this job for all pairs of sentence relations results clustering sentences based on their meanings. Clustering sentences is done by *α-cut* rule [17,18].

Text Tokenizing and Syntax Parsing

For obtaining words as a noun, verb, noun phrase or verb phrase that can extract it's meaning from corpus we need first distinguish it's part of speech via a tokenizer and a syntactic parser based on Persian language grammar. For reaching this goal we need a suitable grammar. As we know a natural language grammar is unrestricted and this matter makes trouble for parsing because of ambiguity and making several parse tree for a sentence. For avoiding this problem a method is selected that converts a natural language grammar to a context free grammar and is not ambiguous, named *two level grammar* which contains some meta variables with initializing them we can obtain a context free grammar based on the value of those meta variables and then this grammar can be parsed much more easier [19]. Of course for this job we need a bulk of rules that initialize the value of these meta-variables and this restriction makes us unable to cover wide area of a language.

Lemmatizing and Stemming

Lemmatization is a function that eliminates the overhead of any word and extracts root or lemma of it. If the root of a word is obtained then finding the meaning of that word becomes much more convenient [12]. Persian's and Arabic's words have four overhead types that includes [9]:

A. Enclitics- objective connected pronouns like BICHAREAM that the lemma is BICHARE (means poor) [6].
B. Suffixes- plural sign or relative adjective signs like BARG HA that BARG is the lemma of it or IRANI that its lemma is IRAN.

C. Proclitics- like AL in Arabic words.

D. Prefixes- that can be noun, adjective or pronouns like HAMANDISHI that its lemma is ANDISHE.

Stemming is concerned with finding the word stem of a given word by removing suffixes from it according to a set of rules [20]. Some of these algorithms used a n-gram based tokenization and this technique solves the problem of rich agglutinative morphology and compounding [10,11].

Knowledge Base Creation for Synonym Words

As we said before the knowledge base for the synonym words is a fuzzy relation.

Our universal set is *W* that is set of all words in the text. These words can be noun, adjective, verb or any phrasal expression those are used in our Persian text. Now we want to obtain words that can be substituted with each other in sentences [5] and for reaching this we need a fuzzy relation between set *W* and itself [14]. We name this relation $\tilde{P}$ the first letter of the word *Paradigmatic*.

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

w_1, w_2 are the words in Persian language and *W* is their set. $\tilde{P}$ is the paradigmatic relation between these words that is also a fuzzy relation. It's membership function is as below:

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

The value of this function is between zero and one based on how much the words w_1 and w_2 are near to each other. Let's make an example. Assume that we have three words in a language means $|w| = 3$ and each of these words are related with another via a membership function and this value express semantic similarity between two words and should be determined by a literature specialist. [Table-1] presents above example for explaining this relation.

Table 1- Fuzzy Relation $\tilde{P}$ for *W* = {gloves, mittens, gardening gloves}

	gloves	mittens	gardening gloves
gloves	1	0.6	0.7
mittens	0.6	1	0.2
gardening gloves	0.7	0.2	1

Distinguishing of Sentences Similarity Relation

At first a fuzzy relation for any sentence should be created. This relation likes a vector that have n components and n= | w |. It means this fuzzy relation relates a sentence with all the words in our knowledgebase. If a word exists in a sentence it's membership function value is 1 and if it doesn't exist the value is 0. Now we should determine which words in the knowledgebase can be substituted with the word in a sentence. For reaching this goal we can compose this sentence relation with the relation that shows our knowledgebase, so any words that could be substituted with it's synonym in the sentence it's membership value is between zero to one. This composition is a fuzzy max-min composition between the sentence relation and the knowledgebase relation named P described in previous section. At this point we have a fuzzy relation for any sentence that shows which words or their synonyms exist in it. Now for determining similarity between these sentences we use a fuzzy proximity relation between the fuzzy relation of the sentences. The name of this relation is fuzzy tolerance relation [15]. This relation must be reflexive and symmetric and if transitive property adds to it, it will be a similarity relation. We define this relation as follows [16]:

If we have a relation between two sets $X = \{x_1, x_2, ...\}, Y = \{y_1, y_2, ...\}$ and fuzzy relation R_{y_i} is a set or subset of *X*'s that relates with y_i and R_{y_j} is a set or subset of *Y*'s that relates with y_j then the similarity between R_{y_i} and R_{y_j} is defined as:

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

as you see if $\tilde{A}$ is a fuzzy set then according to definition, $|\tilde{A}|$ is cardinality of fuzzy set $\tilde{A}$ and it's value is obtaining as follows [13,14]:

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

and here *S* is the cardinality of intersection of R_{y_i} and R_{y_j} divide by minimum of cardinality of one of R_{y_i} or R_{y_j}. The *S* relation defined above is a proximity relation because it is reflexive and symmetric so we can use it for distinguishing the similarity of sentences. We can use from *α-cut* for clustering of sentences those are similar to each other. This is reached via a fuzzy similarity relation like $S \geq S_\alpha$ based on a suitable *α-cut* and this is a very good progress in a multi-document summarizing system.

Results

This system is tested by a text with 58 sentences that contains 15 clusters of the same meaning sentences based on distinguishing of a human specialist. Each cluster have some sentences that have the same meaning and number of these sentences and their normal weights mentions in the table below. System initializes $S_\alpha = 0.7$ and after running on this sample makes 22 clusters of the same meaning sentences based on the knowledgebase that contains 946 words and synonyms. The error rate of the system shows in the [Table-2].

Table 2- 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	22.20%
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	33.30%
C6	8	0.8*1/15	8	0%
C7	9	0.9*1/15	7	22.20%
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%

So if we calculate the average of error rate based on cluster weights as below:

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

We will reach to 7.66% error. This means that system works at rate of 92.34% correctly on this sample.

Discussion

In this approach we found that text can be segmented via a fuzzy proximity relation. The point that is obtained from this research is if the α value in S_α is increased and get near to one then the system error will decrease. But we set S_α to 0.7 because in creating knowledgebase we had error in determining fuzzy membership between words and phrases that increase the error so with setting $S_\alpha = 0.7$ we are trying to delete the effect of that error.

Conclusion

This manner prepares a solution for detecting the same meaning sentences based on paradigmatic relation. It means that if a word substitutes with it's synonym in a sentence, this manner can help distinguishing the similarity and preparing the ability of selecting one of them for inserting in summary in order to avoiding redundancy in it.

Acknowledgment

The author wishes to thank Dr. Mostafa Assi and Dr. Mohammad Reza Kangavari for their helpful supports.

References

[1] Goldstein J., Mittal V. Carbonell J. & Callan J. (2000) *CIKM International Conference of Information and Knowledge Management*, Mclean VA, USA, 165-172.

[2] Poormansoori A., Kahani M., Varaste S. & Kamyar H. (2011) *International Conference on Asian Language Processing*, 145-149.

[3] Shahabi A.S. & Kangavari M. (2002) *Advances in Cognitive Science*, 4(3), 35-41.

[4] Shamsfard M., Akhavan T. & Erfani M. (2009) *World Applied Science Journal,* 7, 199-205.

[5] Aboumahboob A.. (1996) *Farsi Language Structure*, Mitra Pub.

[6] Natel Khanlari P. (1991) *Farsi Language Grammar*, Toos Pub.

[7] Shahabi A.S. (1997) *Farsi Text Understanding*, MS Dissertation.

[8] Bateni M.R. (1992) *Language Grammar a New Look*, Agah Pub.

[9] Dichy J., Krauwer S. & Yaseen M. (2001) *Workshop on Arabic Language Processing: Status and Prospects*, 20-23.

[10] Halacsy P. & Tron V. (2006) *Working Notes for the CLEF Workshop*, 4730, Springer Heidelberg.

[11] Miangah T.M. (2006) *Journal Quantitative Linguistics*, 13(1), 1-16.

[12] Siemens R.G. (1996) *CH Working Papers*, 1(1).

[13] Wang L.X. (1997) *A Course on Fuzzy Systems and Control*, Prentice Hall Inc.

[14] Zimmermann H.J. (1996) *Fuzzy Set Theory and its Application*, 3rd ed., Kluwer Academic Pub.

[15] Dubois D. & Prade H. (1980) *Fuzzy sets and systems Theory and Applications*, Academic Press Inc.

[16] Fujimato T. & Sugano M. (1997) *Sixth IEEE International Conference on Fuzzy Systems*, 1, 231-234.

[17] Marcu D. & Gerber L. (2001) NAACL-2001 *Workshop on Automatic Summarization*, 1-8.

[18] Yang M., Wu K.L., Hsieh J.N. & Yu J. (2008) *IEEE Transactions on Cybernetics,* 38(3), 588-603.

[19] Krulee G.K. (1991) *Computer Processing of Natural Language*, Prentice Hall Inc.

[20] Heinig C. & Mehn F. (2006) *Student Research Workshop on Computer Applications in Linguistics*.