A FUZZY APPROACH FOR PERSIAN TEXT SEGMENTATION BASED ON SEMANTIC SIMILARITY OF SENTENCES

Amir Shahab Shahabi , Dr. Mohammad Reza Kangavari
Islamic Azad University South Tehran Branch,shahabi_amir@azad.ac.ir , Science & Industry of Iran University

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 approximation 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.

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

1. 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 one of the most important factors in Multi-Document Summarization [Goldstein J. , et al . (2000)]. 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 [Natel Khanlari , P. (1991)] , [Aboumahboob , A. (1996)]. 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 [Shahabi , A. Sh. (1997)]. After finding words and nouns and verb phrases by tokenizer and syntactic parser the lemma of words and verbs is created by lemmatizer [Natel Khanlari , P. (1991)] , [Siemens R. G. (1996)] , [Dichy J. , et al. (2001)] , [Bateni , M. R. , (1992)]. 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 [Zimmermann H. J. (1996)], [Wang L. X. (1997)] 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 [Dubois D. et al.(1980)] , [Fujimato T. et al.(1997)]. Repeating this job for all pairs of sentence relations results clustering sentences based on their meanings. Clustering sentences is done by $\alpha - cut$ rule [Marcu D. et al. (2001)].

2. 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 that is not ambiguous, named

two level grammar that 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 [Krulee G. K. (1991)]. 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.

3. LEMMATIZING

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 [Siemens R. G. (1996)]. Persian's and Arabic's words have four overhead types that includes [Dichy J., et al. (2001)]:

1. Enclitics – objective connected pronouns like BICHAREAM that the lemma is BICHARE (means poor) [Natel Khanlari P. (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. Proclitics – like AL in Arabic words.
4. Prefixes – that can be noun, adjective or pronouns like HAMANDISHI that its lemma is ANDISHE.

4. 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 [Aboumahboob , A. (1996)] and for reaching this we need a fuzzy relation between set W and itself [Zimmermann H. J. (1996)]. 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. Its membership function is as below:

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

the value of this function is between zero to 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 sentences with their words as below and each of these words are related with each other via a membership function and this value express semantic similarity between them and should be determined by a literature specialist. Based on these sentences and above assumption we can define our knowledgebase. First of all we state the sentences in English:

- 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.

At these sentences there exists similarity relation in meaning that we intend to find it via this method. Now we should create a knowledgebase of words and synonyms of these sentences. The word and phrase set of our example is as below:

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

and the fuzzy relation that specifies our knowledgebase is as follows:

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

5. 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 its membership function value is 1 and if it doesn't exist the value is 0. For our example the fuzzy relations for each sentence are as follows:

Table 2. Fuzzy Relation of each sentence

		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

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 its synonym in the sentence its membership value is between zero to one. This composition is a fuzzy max-min composition between the sentence relations $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3$ and the knowledgebase relation named $\tilde{P}$ 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. For our example the results of their compositions are as follows:

Table 3. Fuzzy Max-Min Composition between sentences & knowledgebase

		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

Now for determining the similarity between these sentences we use a fuzzy proximity relation between the fuzzy relations of the sentences. The name of this relation is fuzzy tolerance relation [Dubios D. et al. (1998)]. 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 [Fujimato T. et al. (1997)]:

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 below:

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

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 [Wang L. X. (1997)][Zimmermann H. J. (1996)]:

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

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. For our example the fuzzy proximity relation of the example's sentences are as follows:

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

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

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

So the similarity between the first and second sentences is so much but they differ from the third sentence.

We can use $\alpha - cut$ 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 $\alpha - cut$ and this is a very good progress in a multi-document summarizing system.

6. 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 has 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 below:

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%

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.

7. 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.

8. 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.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Mostafa Assi.

REFRENCES

1. Aboumahboob A. 1996. Farsi Language Structure. Mitra Pub.
2. Bateni M. R. 1992. Language Grammar a New Look. Agah Pub.
3. Dichy J., Krauwer S., Yaseen M., "On Lemmatization in Arabic, A formal Definition of Arabic Entries of Multilingual Lexical Databases," Proc. of the workshop on Arabic language Processing: Status and Prospects, PP. 20-30, July 6th , 2001. Association for Computational Linguistics 39th Annual Meeting and 10th Conference of European Chapter , Toulouse.
4. Dubois D., Prade H. 1980. Fuzzy sets and systems Theory and Applications. Academic press Inc.
5. Fujimato T., Sugano M., 1997. "Clustering verb, Adjective, Adjectival verb concepts using Proximity Relation," IEEE.
6. Goldstein J., Mittal V. Carbonell J., Callan J. ,"Creating and Evaluating Multi-Document Sentence Extract Summaries," Proc. of the 2000 CIKM International Conference of Information and Knowledge Management. Mclean VA, USA, PP. 165-172. 2000 November.
7. Krulee G. K. 1991. Computer Processing of Natural Language , Printice Hall Inc.
8. Marcu D. , Gerber L., "An Inquiry in to the Nature of Multi-Document Abstract , Extracts and their Evaluation," Proc. of Automatic Summarization Workshop, 2001.

9. Natel Khanlari P. 1991. Farsi Language Grammar. Toos Pub.
10. Shahabi A. Sh. 1997. Farsi Text Understanding. MS Dissertation.
11. Siemens R. G., "Lemmatization and Parsing with TACT Preprocessing Programs," Department of English University of British Columbia, 1996.
12. Wang L. X. 1997. A Course on Fuzzy Systems and Control. Printice Hall Inc.
13. Zimmermann H. J. 1996. Fuzzy Set Theory and its Application, Third Edition. Kluwer Academic Pub.