An Automatic Linguistics Approach for Persian Document Summarization

Hossein Kamyar, Mohsen Kahani, Mohsen Kamyar, Asef Poormasoomi
Web Technology Lab, Ferdowsi University of Mashhad
Mashhad, Iran
Hossein.kamyar@stu-mail.um.ac.ir,kahani@um.ac.ir,mkamyar@stu-mail.um.ac.ir,As.poormasoomi@stu-mail.um.ac.ir

Abstract **__ In this paper we propose a novel technique for summarizing a text based on the linguistics properties of text elements and semantic chains among them. In most summarization approaches, the major consideration is the statistical properties of text elements such as term frequency. Here we use centering theory which helps us to recognize semantic chains in a text, for proposing a new automatic single document summarization approach. For processing a text by centering theory and extracting a coherent summery, a processing pipeline should be constructed. This pipeline consists of several components such as co-reference resolution, semantic role labeling and POS [Part of speech] tagging.**

Keywords- *Single-document summarization, Centering Theory, LSI, Extractive, Persian*

I. INTRODUCTION

Automatic documents summarization is an important tool in the age of explosive growth of data. According to [1] summary refers to a generated text from one or more texts and it consists of important concepts of the texts. This generated text should not be bigger than half of the source texts. This simple interpretation involves main properties of a summary: (1) summary of one or more texts, (2) major information of the source texts, and (3) short.

Investigations about extracting important and salient knowledge from a text are subject of single document summarization [2]. The researches in this field can be categorized into extractive and abstractive summarization. Extractive summery means returning of some sentences as important sections, and abstractive summary means representation of internal knowledge of a text using possibly different wording [2].

In this work, we propose an extractive single document summarization approach using a combination of a linguistics theory (Centering Theory) and some statistical parameters of text. The proposed method tries to address the current challenges of summarization approaches: (1) Longer length of the extracted sentences than the average length of source sentences, (2) Dispersion of data in the text, (3) Similarity of information between extracted sentences, (4) Lack of coherence in generated summary, (5) Dependence of the summary to the statistical parameters of the text elements such as term frequency and etc. For solving the first problem, we used statistical parameters and for other problems we used the centering theory.

The remainder of the paper is organized as follows: Section 2 discusses related works in single document summarization in English and Persian as well as the literature review on centering theory. In Section 3, we describe the proposed method in details. The experimental results are presented in Section 4, and finally conclusion is drawn and future works are discussed.

II. RELATED WORKS

A. Extractive single document summarization

Many approaches are proposed for single document summarization each of which belong to one of computational text categories such as machine learning, genetic algorithms, neural network, fuzzy, clustering and statistics. On English, in investigation [3], LSI algorithm, as a clustering approach, has been utilized as a logarithmic evidence for term weighting. In [2] with the use of a neural network on DUC2001 dataset, first sentence of each news text as the most important of the sentences is recognized. Also in [4] by using of Centering theory, a summarization method is represented. In this method, CB [Backward looking center] parameter for each sentence is computed and then similar CBs in the whole text are enumerated. Next, sentences that include CB, which belongs to numerous CBs, are selected as important sentences. Article [5] constructs utterance topic model to generating a coherent summary with the utilization of centering theory and LDA [Latent Dirichlet Allocation]. The idea that centering theory can recognize coherence in the text is the major contribution of this paper. This paper focuses on DUC2005 [Document Understanding Conference], TAC2008 [Text Analysis Conference], TAC2009 and it reports good results for summarization.

Unlike English-written text summarization methods, summarization of single and multiple documents written in Persian language is a relatively new field of research.

The first work on Persian Language is FarsiSum in 2004[6]. It is a Web based application programmed in Perl and based on SweSum [7]. FarsiSum selects sentences from documents with the main body of language independent modules implemented in SweSum. It has added the Persian stop-list in Unicode format and has adapted the interface modules to accept Persian texts. The next work was done by Karimi and Shamsfard [8]. It is a Persian single document summarization method based on lexical chains and graph based methods. Zamanifar in [9] proposed an integrated method for Persian text summarization which combines the term co-occurrence property and conceptually related feature of Persian language.

B. Centering Theory

Centering theory [10] is one of the components of general centralization and coherent discourse theory of Grosz and Sidner, which is about local coherence and salience. This theory has been formulated by [11] and is supported by empirical evidences in [12]. Since this theory has good potential for recognizing coherence and

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.52

salience in a text, we have used it as the main idea of this investigation.

Cohesion between sentences (transition) is classified into four categories: Continue, Retain, Smooth-shift and Rough-shift as shown in Table (1). For specifying these transitions, some parameters are added to the theory, giving an algorithmic nature to it. These parameters are:

- CF [Forward looking Center] (U_n): a list of all the references in utterance n ordered according to salience (i.e. grammatical obliqueness)
- CP [Prefer Center] (U_n): most valuable element of CF list that is named preferable center.
- CB(U_n): most valuable noun of utterance n that is realized in CP (U_{n-1}).

Two important applications for centering theory in NLP: anaphora resolution and sentence ordering have been investigated in [13]. Although the applicability of centering theory in Persian has not been evaluated, we used it with English language rules and constraints. The empirical results show good adaption for summarization. However, the application of centering theory in Persian for other purposes, such as anaphora resolution, has to be evaluated.

Table 1. Table of Standard Centering Theory transitions.

	CB(Un) = CB(U_{n-1}) Or CB(U_{n-1}) undef.	CB(Un) ≠ CB(U_{n-1})
CB(Un) = CP(Un)	Continue	Smooth-Shift
CB(Un) ≠ CP(Un)	Retain	Rough-Shift

III. The Proposed Model

The main idea of this algorithm has come from the study on centering theory treatment of cohesion and coherence between sentences of texts. This idea consists of two major parts that coincide on two principle concepts of centering theory: (1) with the use of salience concept, important elements of sentence are specified, (2) by using cohesion concept, sentences with the same important elements as major and cohesive utterances for target summery are chosen. The proposed algorithm consists of two independent phases. In the first phase, preprocessing is done to prepare the text for extraction of important sentences. In the second phase, the important sentences, based on the semantic transitions (centering theory transitions) that are accomplished between utterances, are selected. The proposed algorithm flowchart is shown in Figure (1).

A. *Pre-processing phase*

For processing a text and finding transitions between sentences, we require to traverse some steps that idiomatically are named preprocessing steps. These steps utilized some text processing toolkits implemented by Web Technology Lab of Ferdowsi University of Mashhad [http://wtlab.um.ac.ir]. We used PHP script for generating these toolkits in both online and offline modes. First, the target text should be processed via Co-reference Resolution package. In linguistics, a co-reference occurs when several phrases in some sentences in a text refer to the same phrase. Substituting a pronoun or a noun phrase with noun reference leads to accurate and transparent comprehension and interpretation of the text and obviates the ambiguity of it.

After splitting text to sentences, results of Co-reference Resolution preprocess step, will be used as the input of the SRL (Semantic Role Labeling) step. This tool assigns semantic role of each elements of the sentence with respect to the verb. While carrying out the operation, syntactic role of each element such as subject, indirect object, direct object, adverb and … is recognized. Because some roles may be a phrase, we need to process the text by the POS tagger to specify one indicant for each role group. Then for each text, one matrix is constructed. As shown in (1), the matrix rows consist of sentences and matrix columns contain of syntactic role of each term in a sentence.

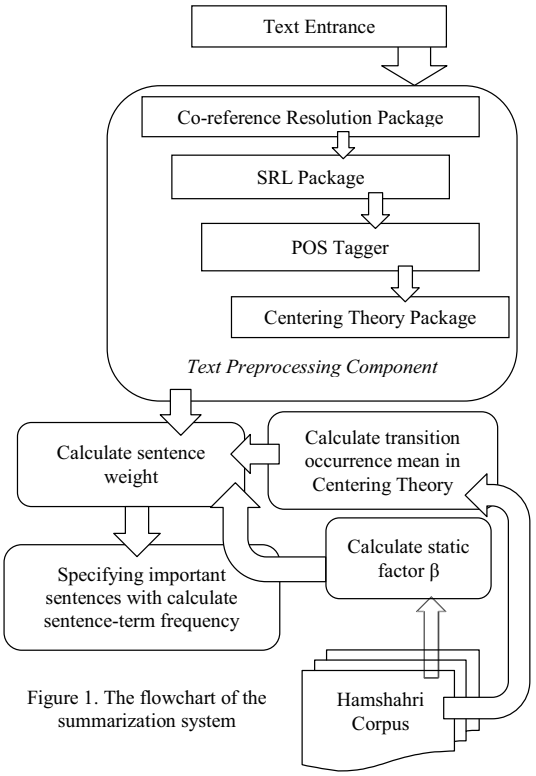

Figure 1. The flowchart of the summarization system

(1) Sentence-Term Matrix:

$$\begin{array}{c} \\ \text{Sentences} \end{array} \begin{array}{c} \text{Terms} \\ \begin{bmatrix} Term11 & \cdots & Term1N \\ \cdots & \cdots & \cdots \\ TermM1 & \cdots & TermMN \end{bmatrix}_{M \times N} \end{array}$$

B. *Specifying the centering theory parameters and semantic transition between sentences*

In this phase we specify the centering theory parameters for each sentence of a text. According to the centering theory definition, the order of nouns in a sentence is as follow:

Subject > Indirect object > Direct object > Other

As mentioned earlier in the sentence-term matrix, columns that show terms are tagged with syntactic roles. Thus up to this step, we have constructed CF and CP parameters for each sentence. Pseudo-code for specifying CB parameter is seen in Figure (2).

```
For sentence number (n)
Tag the most important term as CP
```

```
X = CP(n);
End;
CB(n+1) = NULL;
For each term in sentence number (n+1)
        If(Sentence-Term((n+1),i) = X)
                CB(n+1) = X;
        End;
End;
```

Figure 2. Pseudo-code for calculation of CB parameter

Now we implement centering theory rules for four basic semantic transitions on tagged sentence-term matrix. Pseudo-code for this operation is shown in Figure (3). The new matrix has dimensions of m×(n+1) in which column (n+1) shows semantic transition that occurs between the current and previous sentence.

```
For each sentence
  If(CB(m) = CB(m-1) Or CB(m-1) undef)
    If(CB(m) = CP(m))
       Sentence-Term(m,n+1) = Continue;
    Else Sentence-Term(m,n+1) = Retain;
  If(CB(m) ≠ CB(m-1))
    If(CB(m) = CP(m))
      Sentence-Term(m,n+1) = Smooth-shift;
    Else Sentence-Term(m,n+1) = Rough-shift;
End;
```

Figure 3. Pseudo-code for calculation of transition between Sentence-Term matrix rows

C. Sentences selection phase

To select the important sentences, the following steps should be followed:

1) Sentence weighting based on transition occurred between two successive sentences:

In this work, the important sentences are selected according to a heuristic based on semantic transitions structure; some sentences on the basis of transition type are selected as important sentences. In fact semantic transitions in centering theory show the relation between nouns of two successive sentences. Consequently, we consider each transition alone, and then evaluate selected sentences according to this transition. For each transition, the mean occurrence between sentences in the whole corpus is calculated as shown in Table (2).

Table 2. Table of mean occurrence of transitions in Hamshahri2 corpus

Transition	Occur Average in Hamshahri2
Continue	0.15
Retain	0.24
Smooth-shift	0.19
Rough-shift	0.42

2) Calculating transition weight based on it frequency on the corpus:

Transition weighting formula according to its mean for selecting important sentences is shown in (2). Calculated weight for each transition is a static numerical value that is calculated from Hamshahri2 [http://ece.ut.ac.ir/dbrg/hamshahri] corpus as a learning dataset. Acquired investigations and results show considerable accuracy of calculated factor.

$$TF = \left(\frac{f_{transition,i}}{\sum_k f_{transition,k}}\right) Log_2\left(\frac{|D|}{|\{d:transition \in d\}|}\right) \quad (2)$$

3) Transition weighting based on frequency on a text

According to a heuristic, high frequency rate of a transition between sentences of a text reduces its reliability. For this reason, by calculating relative frequency for each transition in each text and normalizing it, the problem will be solved. Equation (3) shows calculation of this normalized relative frequency.

$$TW_n = \beta \times (MT_n - MA) / VA \times TF \quad (3)$$

Where $VA := {}^1/_4 \sum_{i=1}^4 (MT_i - MA)^2$

And $MT_n := \frac{\sum_{i=1}^{number\ of\ this\ transition\ n} 1}{\sum_{k=1}^{number\ of\ transitions} 1}$

And $MA := {}^1/_4 \sum_{j=1}^4 \frac{\sum_{i=1}^{number\ of\ this\ transition\ j} 1}{\sum_{k=1}^{number\ of\ transitions} 1}$

According to our investigations on learning corpus (Hamshahri2) β factor for each transition is as shown in Table (3). Now we can choose important sentences based on the calculated transitions weight for a text using (3). These sentences are the summery of the text.

Table 3. Table of β factor for each transition calculated by learning on Hamshahri2 corpus

Transition	β Factor
Continue	2.6
Retain	1
Smooth-shift	1.8
Rough-shift	0.85

D. Sentences weighting based on calculates relative frequency of term-sentence and normalized it

The dimensions of selected sentences may be more than the number of expected sentences for the summery. It can happen because transition types that occur between sentences may be the same and all or major parts of sentences are selected. If this situation occurs, another step has to be done to reduce the summary size. In this step we combine sentence weights based on the transition with sentence weight based on its terms.

1) Terms weighting:

For each term in a sentence according to (4) amount of TF-ISF is calculated.

$$(TF \times ISF)_{j,i} = \left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) Log_2\left(\frac{|D|}{|\{d:t_i \in d\}|}\right) \quad (4)$$

In (4) $tf_{j,i}$ is the value of the term frequency in the sentence, $|D|$ is the total number of sentences of the text, $|\{d: t_i \in d\}|$ is the number of total sentences that term i is realized in them. Now, using (5), we calculate mean of terms weight for a sentence and normalize it. In fact this value is a new sentence weight.

$$SMT_j := \frac{\sum_{i=1}^n NTFPS_{j,i}}{\sum_{i=1}^n NTFPS_{j,i}>0} \quad (5)$$

Where $NTFPS_{j,i} := \left(\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) - MTPS\right) /VPS$

And $MTPS_j := \frac{\sum_{i=1}^n \left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right)}{\sum_{i=1}^n tf_{j,i}>0}$

And $VPS_j := \frac{\sum_{i=1}^n \left(\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) - \sum_{i=1}^n \left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right)\right)^2}{\sum_{i=1}^n tf_{j,i}>0}$

Now we multiply the weight value for each sentence in the obtained weight from (5) and then enter its result into the column *n+2* of the sentence-term matrix. This matrix is shown in (6). Then we sort the sentence-term matrix based on the column n+2 and select the sentences according to the density rate of the expected summery.

(6)

Sentences \ Terms						
	$Term11$	$\cdots$	$\cdots$	$\cdots$	$Term1N$	$Transition1$	$Weight1$
	$\vdots$	$\ddots$	$\cdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	$TermM1$	$\cdots$	$\cdots$	$\cdots$	$TermMN$	$TransitionM$	$WeightM$

$M\times(N+2)$

IV. EVALUATION

In this section we describe the data set used for the evaluation, the implemented systems and the experimental results.

1) Dataset : unfortunately there is not standard dataset such as DUC data set for text summarization in Persian language. There are several NLP Persian corpuses exist, such as Bijankhan corpus for POS tagging, Hamshahri1/2 corpus for text retrieval, TEP corpus for translation and etc. In this work we used Hamshahri2 corpus to perform our experiments. This corpus contains about 4000 documents in TDT TREC format, each document belongs to a subject. These documents were gathered from Hamshahri newspaper between years 1996 to 2007. Here, the cluster 2007 has been used and the text of the news have been extracted from the corpus using a XML parser. Then for each document, we generate four human abstractive summeries which summary methods are comprised with them.

2) Evaluation tool: As there is no standard tools to do comparison between summeries in Persian, we decide to evaluate our method in two ways. *a)* First, we implement ROUGE [Recall Oriented Understuy for Gisting Evaluation] [14] evaluation tools for Persian language. ROUGE is the most commonly used tool for text summarization task in English. Each method estimates recall, precision and f-measure between human written reference summaries and the candidate summaries of the proposed system. For example in ROUGE-N, the n-gram recall is computed as shown in (7):

$$ROUGE - N = \frac{\sum_{S\epsilon\,\{Reference\ Summaries\}} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S\epsilon\,\{Reference\ Summaries\}} \sum_{gram_n} Count(gram_n)} \quad (7)$$

where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the candidate summary and a set of reference summaries. *b)* we asked some master students from the linguistics department to evaluate the summaries by voting between 1(very poor) and 5(very good).

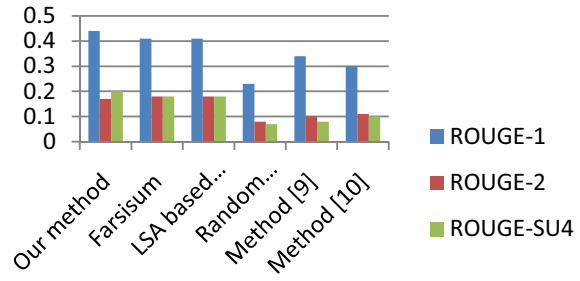

Figure 4. Performance results between methods and Human summaries by implemented ROUGE

3) Impelementing the summarization systems: In order to evaluate our proposed method properly, we compare our method by the only existent system in Persian, FarsiSum. In addition we implemented the following systems:

- LSA base summarizer proposed by Gong [3].
- Random base summarizer.
- Our proposed method
- Method proposed in [8]
- Method proposed in [9]

4) The experimental results: we consider F-measure from ROUGE-1, ROUGE-2 and ROUGE-SU4 comparison results which are shown in Figure (4). In Figure (4) the average results for 90 documents in Hamshari2-2007 for all of approaches. We see F-measure parameter in (8):

$$\text{F-measure} = 2 \times \frac{precision \times recall}{precision + recall} \quad (8)$$

V. CONCLUSION

In this paper, in regard to existence extractive single summarization problems, a new method by combining a lingual theory and statistics text parameters is proposed. Our experimental results show that this work is better than other implemented approaches in Persian. Some of the attractive fields about centering theory applications in Asian languages are centering theory evaluation for Persian, Arabic and the other, concept extraction based on Centering theory, multi document summarization with extract semantic transition pattern and etc.

Acknowledgment

This work was supported by the Web Technology Laboratory of Ferdowsi University of Mashhad. We would like to thank WTLAB group members.

REFERENCES

[1] D. R. Radev, E. Hovy, K. McKeown, "Introduction to the special issue on summarization," Computational Linguistics., vol. 28, Issue. 4, 2002, pp. 399-408.

[2] D. Das , A. Martins, "A Survey on Automatic Text Summarization," Literature Survey for the Language and Statistics II Course at CMU, 2007.

[3] Y. Gong, X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," Proceedings of the 24th annual international ACM SIGIR conference., New Orleans., Louisiana.,United States., 2001.

[4] L. Hasler, "An investigation into the use of centering transitions for summarisation," In Proceedings of the 7th Annual CLUK Research Colloquium., University of Birmingham ., 2004, pp. 100—107.

[5] D. Pradipto, R. Srihari, "Utterance Topic Model for Generating Coherent Summaries," In Proceeding of TAC(NIST)., 2009.

[6] N. Mazdak, "FarsiSum-a Persian text summarizer," Master thesis., Department of linguistics., Stockholm University., 2004.

[7] H. Dalianis, "SweSum A Text Summarizer for Swedish," Technical report., TRITANA-P0015., IPLab-174., 2000.

[8] Z. Karimi, M. Shamsfard, "Summarization of Persian texts," In Proceedings of 11th International CSI computer Conference., Tehran., Iran., 2006.

[9] A. Zamanifar, B. Minaei, M. Sharifi, "A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of the Text," Ninth ACIS International Conference., DOI 10.1109/SNPD., 2008.

[10] A. K. Joshi, S. Weinstein, "Control of inference: Role of some aspects of discourse structure–centering," In Proc. International Joint Conference on Artificial Intelligence., 1981, pp. 435– 439.

[11] W. Chafe, "Givenness,contrastiveness, efiniteness, subjects, and topics," C. Li, editor, Subject and Topic. New York: Academic Press, 1976, pp. 25–76.

[12] W. Kintsch, T. van Dijk, "Towards a model of discourse comprehension and production," Psychological Review., 1978, pp. 363–394.

[13] T. Milan, "Extending Centering Theory for the Measure of Entity Coherence," MSc thesis, Simon Fraser University, 2009.

[14] C. –Y. Lin, Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," In *Proccedings of NLT-NAACL.*, 2003.