USING MACHINE LEARNING METHODS TO SUMMARIZE PERSIAN TEXTS

MEHRNOOSH BAZRFKAN\textsuperscript{a1} AND MUOSA RADMANESH\textsuperscript{b}

\textsuperscript{ab}Department of Computer, Islamic Azad University, Science and Research Branch-Bushehr, Bushehr, Iran

ABSTRACT

Automatic text summarization is a process to reduce the size of text document using a computer program in order to create a summary of text retaining key points in document. Since information keeps increasing and data has been increased, automatic text summarization is practical method desired. This method helps to summarize information contained in a text file with preserving the original contents in the meaning. Summarize large documents is strongly difficulty. One of the problems existing today is the very websites arisen rapid growth and high volume data whereby this can increase the necessity of strong and complicated summarization to reduce data volume and increase speed of access. Technology has provided this possibility to build a cohesive summary given the Length, style of writing and writing style. For instance, Google search engine uses this possibility.

KEYWORDS: Machine Learning, Text Summarization, Support Vector Machine Classifier

The most important factors arising emergence and expansion of Knowledge of text summarization grown in recent years have been increasing trend of knowledge and information across the world, the dramatic expansion of the Internet, communication and information exchange as well as the substantial advancements in Memory Stick technology. These all have caused the users face problems in categorizing the text information including documents, articles and electronic letters and etc whereby needing to expand the automatic classification of text has been increased throughout the world (Joachims, 1996) (Dumais et al. 1998)(Pavlov et al. 2001). Database Management, Natural Language Processing, Machine Learning and pattern recognition mentioned as the most important knowledge types which engage with issue of text mining(Dumais et al. 1998)(De Vel et al. 2002). In one of the earliest research on Text Classification in 1961, according to the content of title, the texts can be classified in different groups with different titles and contents regarding their selected key terms (Maron, 1961). In 1988, Hayes in a cooperation with his colleagues, proposed commercial applications of natural language processing techniques in classification problem reports and stories within numerous categories in which Full grammatical and conceptual analysis used instead of Pattern matching techniques and its accuracy was a bit less than human identification(Philip et al 1988). Further, in 1992, Jacobs used some of natural language processing techniques on Knowledge Base to statistical methods so as to provide categories of news, concluded that, however statistical methods simplify performing and need just a little manipulation, adding pre-processing methods to natural language processing techniques increase accuracy and pace of work(Paul and Jacobs, 1992). Gvttrand colleagues (1993) created a system to categorize electronic letters using artificial neural network, provided with a dictionary and let having ability to learn. Classification using this method with accuracy to the extent of 79.1% was close to human’s accuracy with 79.4 % (Geutner et al, 1993). In 2005, Saucy and Mineau used k-Nearest Neighbors algorithm for reducing the features extracted in text classification. In this study, use of Machine Learning Methods to Summarize Persian Texts has been studied where support vector machine classifier of Machine Learning Methods used to Summarize Persian Texts. In recent years, support vector machine in most of Classification issues including Text classification, Image Recognition and so on has been applied successfully (Bolotand Crook, 2008).Support vector machine is a non-probabilistic binary classifier which uses data mapping from the main entry space to a space upward for separation. The model seeks a hyper plane that its distance with data of two classes is maximum. In this method, it is intended to implement a System with a minimum capacity, or, in a plain language, a System with minimum complexity so as to achieve boundary of classes. As a result, support vector machine using lower educational data compared to other methods can estimate the boundaries of system with a proper accuracy without distorting or any mistake in extrapolating the system. Indeed, support vector machine classifier using lower educational data compared to other methods, aims to estimate the boundaries of system with a proper accuracy without distorting or any mistake in extrapolating the system. To sum up, support vector machine classifier aims to achieve function f(x), which

\textsuperscript{1}Corresponding author
determines the hyper plane. This hyper plane separates two classes from each other, shown in figure 1. According to this figure, M is a margin, i.e. Hyper plane distance from the nearest points of the two classes (Cyamyng, 2009).

Figure 1: Support vector machines in Linear condition (Cyamyng, 2009)

TEXT SUMMARIZATION

The process of text summarization is divided into three stages: 1-a structured reconstruction of the main text is developed in preprocessing stage, 2-an algorithm needs to transform a text structure to a summarized structure at processing stage, 3-the final summarization can be achieved in producing stage using summarized structure(Jones, 1999). 1-Superficial methodsthatare limited to Syntactic reconstruction and try to be included of important parts extracted from the text, 2- Deeper methods which address the Semantic level of reconstruction of the original text, including some parts of Linguistic processing.

According to the first method, the preprocessing stage is the very Reducing the size of the presentation space, including the factors as follows on the whole:
1-common removal which does not imply semantic content, and not causing data collection, 2- changing all the words written with Capital letters or Lowercase letters, 3- separate similar words and access to their root which focuses on Meaning.

A model used currently is a vector model (Maron, M. E.1961). After processing stage, each text element is taken as an-dimensional vector. Hence, use of Measurement systems for measuring the similarity between elements of the text would be possible. Measurement system used the most is the Cosine measure, defined as \[ \cos \theta = \frac{(x,y)/(|x|,|y|)} . \] \( \langle x, y \rangle \) Shows Scalar product, and \(|x|\) shows UnitX. Hence, Maximumsimilarity achieves in \( \cos \theta = 1\) and \( \cos \theta = 0\) shows the final difference between text elements.

Quality assessment of summarized text in research grounded on summarizing is an important issue. In an attempt to standard the text summarization methods, an accurate assessment in text summarization evaluation conference was proposed (Jones, 1999). A series of reference summaries written by reviewers collected to be compared with the summaries from the systems participating in the conference. Another problem reported is that reviewers do not agree on this point (Mani, 2001). More importantly, the summaries proposed by a reviewer at two different time intervals have a long in common only to the extent of 55% (Mani et al. 1998). Idea of reference summary is very important because if such thing exists, then performance of production methods for automatic summary using classical information retrieval can be evaluated practically. In this case, a term is called true while belonging to the reference summary. About the summaries with the specific extent, both scales are equal because the reference summary and the summary extracted are both equal.

Mani and Bloedorn(1998) have proposed an automatic method to produce reference summary. If the main text includes a summary provided by author, then the summary of reference extracted would be consisted of the terms with the most points in common, regarding the Cosine measure provided as the summary by author. In this regards, the method of access to reference summaries event for the texts in large scale would be facilitated. If a
series of texts and reference summaries provided, then
machine learning approach can be used (Salton and
Buckley, 1988). Using an algorithm in machine learning
approach helps to achieve a summary of a series of texts
and documents. In this regards, the terms used in each text
are modeled as the Features of vectors extracted from the
text.

Summarization operations is a classification of
two categories where a term belongs to the Summary
extracted from reference, it would be taken true,
otherwise wrong. It is expected that Summary of
Significant Learning Patterns which results in a summary
can be learnt with diagnosis of the features associated to
the accurate classifications. While a new text is given to a
system, the patterns learnt are used to classify the terms in
accurate and inaccurate classifications whereby a
summary is produced. The important issue lies in a
framework that how to diagnose the features associated to
each other.

Machine learning methods

In the 1990s, with emergence of Machine
learning methods in NLP, a series of seminal publications
using statistical methods to build summaries of document,
emerged. While, in beginning, the feature of
independence assumed the most in the systems, relying on
naïve-Bayes and other systems focused on proper features
without an independent hypothesis. Other important
approaches include hidden Markov models and log-linear
models so as to improve extractive summarization. In
contrast, recently a paper has used neural networks and
third part features so as to improve extractive single-
document summarization. In following, all these
approaches are defined in details.

Naïve-Bayes Simplification methods

Kupiec et al.(1995) proposed a method driven
from Edmundson(1969), able to learn through data.
Classification function of each term using naïve-Bayes
classifier is classified to two groups, valuable in terms of
extraction and invaluable in terms of extract. A specific
term is considered as s. S is a series of terms which
develops a summary and \[ F_1, \ldots, F_k \] are the features. The
independence of the features are assumed as follows:

\[
P(s \in S | F_1, F_2, \ldots, F_k) = \frac{\prod_{i=1}^{k} P(F_i | s \in S).P(s \in S)}{\prod_{i=1}^{k} P(F_i)}
\]

These features were inconsistent with
Edmundson theory (1969), but, sentence length and
uppercase words were considered as well. One score was
given to each sentence regarding the equation, and only
one top sentences were extracted. To evaluate this system,
a series of technical documents using manual was used.
For each sentence in the manual, the authors examined its
Harmony with the sentence with real documents,
proposed a pattern. Thereafter, auto extracts were
analyzed. Analysis of features showed that a system can
have the best performance only while the key features and
positions used together with sentence length. A one et
al.(1999) merged a Navie-Bayes classifier, but with more
important features. They defined a system “Dim Sun”,
shown some features including term frequency (tf) and
inverse document frequency (idf) to extract key words.
Inverse document frequency (idf) calculated of a big
series of the same scope like the documents. Statistically
derived two-noun word collocations were used as the
units for Enumeration together with words. A label”
entity” was used and each entity taken as a token. They
further used some shallow discourse analysis like
reference to the same entities in a text sustained on
cohesion. The references at very shallow level were
resolved by linking name aliases like “U.S” to “United
States”, or ”IBM” to “international business machines”. Synonyms
and morphological variants were merged
together while taken as lexical terms and the previous
lexical tem detected using Wordnet (Miller, 1995). The
companies used in experiments were all from newswire
where some belonged to TREC evaluations.

Rich features and decision making trees

Lin and Hovy(1997) addressed examination of
importance of sentence position. Only weighing of a
sentence with its position in the text, implied as position
method, arise from the fact that, texts follow a verbal
structure, and sentences are developed with the unity of
overall topics in the certain specifiable locations. Yet,
ever since verbal structure changes significantly in Length
Domains, position method cannot be defined as defined in
Baxendale(1958). Position method cannot be simply
defined. A research by Lin and Hovy(1997) examined
tailoring methods of position method to achieve the
optimal status in this method and get to know how this
method can evaluate the effectiveness extent. a series of
newswire used to produce a series of Ziff-Davis texts of
TIPSTER program including a text about computer and
hardware together with a series of key words and a small
summary of 6 sentences. For each document in the series, the authors evaluated the position of each sentence to the key words. Thereafter, the positions of sentences with their performance mean to build optimal position policy were graded. Two types of evaluations conducted. Accordingly, an unseen text to measure whether the same method used in the different context was used. The early evaluation shows contours which were the similar to the educational documents. In second evaluation, word overlap for manual with the extracted sentences was evaluated. Windows in manuals compared with windows assigned to selective sentences and the accuracies were analyzed. A high extent of coverage shows effectiveness of position method. In another study by Lin(1999), assumption by having independent features goes beyond, attempted to model the issue of extracting sentences using decision making trees instead of naïve-Bayes classifier. Most of the features of naïve-bayes classifier were evaluated and its impact on extracting sentence was analyzed as well. Data used in this study found as publicly available collection of texts which were divided to two different topics where on these texts achieved of TIPSTER-SUMMAC evaluations associated to Information retrieval systems. A series of data includes essential text fragments, mentioned essential to bring them in summaries so far as giving response to TREC topics. These fragments evaluated each with human judgment. The experiments defined in this study were conducted using SUMMARIST system within University of Southern California. This system extracted sentences from documents and coordinated them with manual extractions by human like most early studies about extractive summary. Some new characteristics include query signature, IR signature, Numerical data, Proper Name, Determiner, weekday or month, quotation. It should be noted that some characteristics including query signature despite an extrapolated summarization framework, due to the context of evaluation, are grounded on questions. Author has experienced a variety of baselines, like mere use of the situational characteristics and Simple combination of characteristics that add values. While evaluating by means of coordinating extractive sentences by human and machine, decision making classifier would be superior, but, about three topics a simple combination of characteristics would be prevailed. Lin supposed that this issue due to independence of several characteristics would occur.

Hidden Markov models

In spite of previous approaches, mainly grounded on characteristics without ordering, Conroy and O’leary(2001) addressed modeling extraction of a sentence from a document using Hidden Morkov models. The reason to use Sequential model lies in using local dependencies among sentences. Only three characteristics were used: sentence position in a document, the number of terms used in sentence, and similarity of sentence indicating the terms used in the document.

Figure 2: Markove model to extract three summarized sentences from a document (Conroy and O LEARY, 2001)

Hidden Markov models were built in this way: this model includes $2s+1$ Status varied in range from $s$ summary to $s+1$ summary. Authors were allowed to hesitation in non summary states and to skipping next state in summary states. Figure 1 shows a form of Hidden Morkov models with 7 nodes associated to $s+3$. Using TREC data set as the education sets, the authors are required to achieve maximum-likelihood estimation for each Transition probability so far as estimation of $\bar{M}$ would be formed at which elements $(i,j)$ show the Empirical probability of transmission from $i$ to $j$ state. About each state, $i$ was output data, $b(O_i = Pr (O/statei))$, where $j$ is an observed vector of features. They build a simplifying assumption with multivariate normal characteristics. Thereafter, the output function was estimated for each state by means of educational data to calculate the Probability of maximum mean and Covariance matrix. $2s+1$ means were evaluated, but assumed that all the output functions have a common
Covariance matrix. Evaluation by means of comparing with the extractions by humans conducted.

**Log-Linear models**

Osborne (2002) stated that the existing approaches for summarization have been supposed independent from the characteristics. Log-linear models have been used to settle this hypothesis, shown that this system rather than naïve-bayes model produces better extraction. C is assumed as a label, s is an item used for labeling, $f_i$ shown $i^{th}$ characteristic, $\lambda_i$ is the weight for each characteristic. Log-linear model used by Osborne regarding the following:

$$ P(c|s) = \frac{1}{Z(s)} \exp(\sum \lambda_i f_i(c, s)) $$

$$ Z(s) = \sum_c \exp (\sum \lambda_i f_i(c, s)) $$

Where, only two labels can be used: a sentence is extracted or not extracted. Weights are taught in conjugate gradient descent. Authors previously proposed a non-uniform model, claimed a log-linear model for inclusion in a summary tends to reject most sentences. The same probability added to Naïve-Bayes model for comparison. The classification would be as follows:

$$ \text{lable}(s) = \arg \max_{c \in C} P(c|s), P(c) = \arg \max_{c \in C} (\log P(c) + \sum \lambda_i f_i(c, s)) $$

Authors optimized the former probability using f2 classifier as target function in a part of data set. The summaries were evaluated via score f2, where on $f_2 = \frac{2pr}{p+r}$ mentioned as precision and recall criteria to extractions. Generated from human were evaluated. The characteristics including word pairs, sentence length, sentence position and naïve discourse features accounted as internal introduction or internal conclusion. According to the score f2, log-linear model with former probability to naïve-Bayes classifier outperformed.

**Neural networks and third party features**

In 2001-2002, DUC was an idea to build a summary with 100 words from a single news. Nevertheless, addressing statistical importance, the best evaluation systems failed to act better than baseline. This baseline has been analyzed by Nenkova (2005), which considered three machine high lights. Human generated story high lights were not verbatim extractions from the article. The authors evaluated their systems using two metrics: the first metric: three system generated story high lights connected to each other and three human generated story high lights connected to each other and these two blocks compared with each other. The second metric considered ordering and compared the sentences in a unique state. Svore et al. (2007) taught a model of labels and characteristics for each sentence of an article, managed to assume accurate sentence ranking in a test document. This ranking conducted using Rank Net (Burges et al. 2005) where on a pair-based neural network algorithm designed to rank a series of inputs, using the descent gradient method for training. For training set, to score the similarities existing in human generated high lights and a sentence in a document, ROUGE-1 was used. These scores used similarly as soft labels during training. Some of the characteristics used observed based on position or n-grams frequencies. Nevertheless, innovation aspect of framework relates to use of characteristics where extracted data from query logs and Microsoft’s news search engine. Authors have measured that if a sentence includes key words used in Microsoft’s news search engine, there would be more chance to find them in high light. The extractions were evaluated using ROUGE1 and ROUGE2, and the statistical improvements shown in all around the baseline of three first sentences in a document.

**Analysis**

**Extraction methods**

TF-IDF method: TF-IDF is repeated expressions and Inverted repeat of document, showing how much a word is important. In other words, this method provides Numerical statistics, mentioned as the importance of a word in a document or a set of documents. This criterion
is used in weighting method, whereby the important words can be achieved then so that the most prevailing words can then be found.

The method based on clustering: difference topics have been addressed in documents where they have been placed one by one organized in a document, and frequently the important parts have been fragmented to smaller parts explicitly or implicitly. Yet, organizing can be considered as a summary. Some summary makers act based on classification and clustering. Given this method in multi-document summarization, each document can be summarized based on its topics and then summaries can be adapted with each other based on the topics, and finally the final summarization can be created. To build the final document in the clustering state, TF-IDF method is used once more and then the words associated together are checked so that the clusters with more weights are considered as the more important topic and entered to output (the final summary). Document sentences, relying on the characteristics of summarization are classified into two groups of sentences existing in final summary and worthless Sentences. In this study, a new method to summarize Persian texts is proposed. The algorithm shown follow, regarding the first step that indicates each word in a given text has to be separated, then sums included in database of the words required omitting, mentioned the words including who, been, hi, more, if and etc. have to be removed from sentences. The repeated words in texts are necessary to be removed so that each word needs to be repeated once. The set obtained develops valuable words where the Importance of any word is the same times the word repeated in a text, however a word repeats several times in a text, importance would be reported more. The second step mentioned as detection of text sentences. In this part, the characters existing in a sentence needs to be examined so far as face a character, whereby all the characters before point are considered as a sentence and then analysis on a sentence can be conducted at the next steps.

The third step, a vector with the dimension of the number of key words for each sentence is developed at this part of algorithm where on a weight is considered for an element. Hence, a matrix develops by the number of rows number of sentences and column number of keywords. Fourth step, the classification of sentences using k-means algorithm is conducted so that two classifications of sentence are considered.

**CONCLUSION**

This method based on the indices including sentence length, similar topics, and similar keywords act as the input of fuzzy systems. Thereafter, the entire rules are placed in knowledge base for summarization. After this step, a value ranging from 0 to 1 is given to each sentence based on the existing rules and characteristics in Knowledge Base. The value obtained for each sentence indicates the importance of any sentence to place it in the final summary. There exist three membership functions, L, VL (low and very low), M (average), H (high), VH (very high), which the valuation relies on these functions.

**Summarization method based on fuzzy logic**

This method based on the indices including sentence length, similar topics, and similar keywords act as the input of fuzzy systems. Thereafter, the entire rules are placed in knowledge base for summarization. After this step, a value ranging from 0 to 1 is given to each sentence based on the existing rules and characteristics in Knowledge Base. The value obtained for each sentence indicates the importance of any sentence to place it in the final summary. There exist three membership functions, L, VL (low and very low), M (average), H (high), VH (very high), which the valuation relies on these functions.

SVD matrix: singular value decomposition (SVD) is a powerful mathematical tool, which can obtain orthogonal dimensions of multidimensional data. This method uses different logics including KL (a change in image processing), PCA (Principal component analysis and signal processing), LSA (Latent semantic analysis of text processing). According to LSA, SVD acts on a matrix of words developed from the relationship in documents associated together in terms of semantic content. The words used generally in similar topics are those associated to the space. In this method, the words used in topics and content of sentences are extracted from the main document. According to LSA vector, a conceptual relationship in words can be achieved, and while having the relationships the accurate design of methods can help go forward the conceptual deduction on relationship. Ever since SVD can search orthogonal dimensions in sentences, a sentence which represents an assurance...
related to a document can be chosen. This orthogonality assures lack of redundancy.

**Sentence length**

The number of words associated to each sentence with other sentences is in a way shown below:

As observed in the method discussed previously, the first step in summarization is the summarization based on topics proposed in document. Graph theory helps to detect these topics, outperforms. Followed by common steps, the selected sentences are shown as nodes of a graph without display orientation. There exists one node for each sentence. This relationship in graph shows two separate parts: 1- Segmenting topics separately, 2-place similar sentences in one part. This graph lets selection of sentence has a good coverage, having a thorough summary of topics. Having this graph helps to have specific queries. Hence, a summary can be produced from a sub-graph, or a representative can be chosen from each sub-graph for general queries, entered to output. One of the other advantages of this graph, that is, it can be used to detect important sentences. A node having high cardinality, definitely is an important sentence which can be placed in a summary. Graph theory can be used to adapt documents with each other.

The number of key words in each sentence is as follows:

The important part in designing fuzzy system lies in selecting fuzzy rules and membership functions. A set of fuzzy rules and membership functions directly affect fuzzy system performance so that fuzzy system consists of four parts:

1- fuzzifier: the inputs are translated given Linguistic values and membership functions.
2- The inference engine: after fuzzifier, the inference engine refers to if-then rules in order to extract Linguistic values.
3- defuzzifier: at this stage, the variables after exited from extraction, transform to their final value using membership function.
4- Knowledge base

**REFERENCES**


