Original Article

Automatic Text Summarization for Urdu Roman Language by Using Fuzzy Logic

Zeshan Ali1*, Jawad Ali1, Zhang lining1, Zeeshan1, Niaz Muhammad1

School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China

ABSTRACT

In the new era of technology, there is the redundancy of information in the internet world, which gives a hard time for users to contain the willed outcome it, to crack this hardship we need an automated process that riddle and search the obtained facts. Text summarization is one of the normal methods to solve problems. The target of the single document epitome is to raise the possibilities of data. we have worked mostly on extractive stationed text summarization. Sentence scoring is the method usually used for extractive text summarization. In this paper, we built an Urdu Roman Language Dataset which has thirty thousand articles. We follow the Fuzzy good judgment technique to clear up the hassle of text summarization. The fuzzy logic approach model delivers Fuzzy rules which have uncertain property weight and produce an acceptable outline. Our approach is to use Cosine similarity with Fuzzy logic to suppress the extra data from the summary to boost the proposed work. We used the standard Testing Method for Fuzzy Logic Urdu Roman Text Summarization and then compared our Machine-generated summary with the help of ROUGE and BLEU Score Method. The result shows that the Fuzzy Logic approach is better than the preceding avenue by a meaningful edge.

Keywords: Urdu Roman; Fuzzy Logic; Cosine Similarity; Big Data; Machine Learning

1. Introduction

At present, distinctive network stages have overfilled the clients with excessive facts from several resources and areas, advent it mainly tough for them to obtain the ones which they imply to. Besides, tremendous offers of unlimited information over the internet are in the system of natural language text. Hence, the mission of automatically shortening articles and documents into a dense, ”fast-to-grasp” model will become tremendously critical inside the age of data.

Classically, there are techniques for automatic text summarization, that is extractive and abstractive. Extractive strategies simply stumble on the sentences and phrases of the file which comprise necessary, extremely useful information related to the primary subject matter discussed inside the content. Abstractive summarization, but, is a way to make human-like summarization that comprehends selection, reordering, and summarizing the sentences of a file. Although collectedly simple, extractive methods are the generally used tactics due to their reasonable performance and applicability. 

Albeit collectedly simple, extractive approaches are the approaches that are commonly used because of their applicability and satisfactory performance. But on the other side, extractive approaches are not favorable because their performance is much different from the methods used by human beings to condense and summarize different documents and
articles. Similarly, a quality summarization needs rephrasing and concatenation, which is expecting in the summarization of some sort of complex articles or complex news. But on the other hand, abstractive summarization is very complicated for machines to perform the actions automatically and solely. Specifically, the majority of abstractive approaches used the approaches which are supervised deep learning, which certainly makes them inflexible and poor health. To implemented it extra well known, complex-case articles which explain many areas are the summarization of multi-articles, but here it should be noted they are very few in numbers. That’s why the majority of the current applicable approaches which are like summarization have remained extractive merely than the abstractive which are human-like. In this regard, this paper presents a type of model that will benefit the approaches which are based on the state of the art, which can be equally beneficial for abstractive and summarization approaches, which will improve the quality of the summarization which is automatic text in nature. Here we first introduce an extractive model by implementing a fuzzy logic system and is equipped with customized features that are hand-crafted to extract important sentences from the text. This feature by which this model is customized will enhance flexibility, and as a result applicability as well. Secondly, this model improves and activities sequence to sequence network as well. Which includes a decoder and encoder with a mechanism which is based on the context which is encoding, where the word distribution is to abstract those sentences which are to be an abstractive summary. Figure 1 displays the construction of fuzzy logic for extractive text summarization.

![Figure 1. Fuzzy logic system architecture for extractive text summarization.](image)

### 2. Related Works

The improvement of the latest strategies is permanently required to support to clear up problems[^2^-^17^]. Some unique performances have been planned for automatic text summarization that deeds a spread of different strategies. Most of those techniques are extractive text summarization approaches. The modes proposed in[^18^] use low functions for text unit scoring and deciding text units that have the very best scores as summarization. Proposed models in[^19^] use techniques to immediately or use a changed Version of it. Some text weight scoring continuously In-textual content unit scoring is primarily based on statistics with the aid of manner of machine learning techniques for automatic textual content summarization. The discourse shape version is proposed in[^20^] to score sentences.

However, the above-mentioned techniques conflict with the redundancy trouble That leads to a low-satisfactory precision. The first manner that gives with this problem is introduced in[^21^] and is known as MMR (Maximum Marginal Relevance). Many text summarization strategies take advantage of MMR summarization strategies are developed base on the fuzzy
Automatic text summarization for Urdu Roman language by using fuzzy logic

logic method. To get a terrific result and improve the quality of summary, various techniques make the most a mixture of or extra one-of-a-kind technique\cite{22,23,24}. In \cite{25} Writer proposes a model that advantages the gain of range-based techniques to pick out the most numerous sentences, and additionally the advantage of the non-range method which makes use of fuzzy logic and swarm’s intelligence to select the most vital sentences for textual content summarization. Since currently many new and effective system learning strategies are developed that are specifically based on deep gaining knowledge of techniques and a few textual content summarization techniques have been proposed to make abstractive summaries that benefit new gadgets, analyzing fashions. In \cite{26,27} writers recommend a version that uses an encoder-decoder method to research the representation of sentences by using the encoder and to categorize each sentence with the aid of decoder based totally on encoder representations the use of an interesting approach. The proposed version in \cite{28} has components. The first component is a single series version and not using a decoder to gain knowledge of extractive. The second portion has a decoder this is abstractedly trained to generate sentence-extraction probabilities. sequence-to-

3. Data Collection

In this section, we collect different articles from News Website. These articles are present in the Urdu language. we manually convert all these articles from the Urdu Language to the Urdu roman language with the assist of the Online Dictionary. which is showing below in Table 1.

<table>
<thead>
<tr>
<th>Urdu</th>
<th>Urdu Roman</th>
</tr>
</thead>
<tbody>
<tr>
<td>اکہ خان نہا جوہتوں بچھ ریا نہا اکہ ادیم اس کا پیس</td>
<td>ake khan thaawo jotay beach raha tha. ake adme os kai pass aaya aur pooh a khan es jotay ki kia ye qemat hay khan nay Kaha ka 500 ropey. os aadme nay Kaha khan tum bahoot Zaida kha rahay ho to khan nay kaha ka acha tum kitnay datay ho to os aadme nay Kaha ka 100 ropey. khan nay kaha nahin yara tum bahoot kam day rahay ho, jab wo aadme Janay laga to khan nay kaha ka acha eder aao do 100 ropey. os aadme nay khan ko 100 ropey day dieay aur jotay lay liay pher khan nay kaha nay kaha kuda ka Banda abb bhe muj ko 50 ropey manafa hooa ha akstan ke Wazeer e Azam Imran Khan ka Amrici ahbhar Washington post mein shaya honay waly column mein kehna Hai ke Afghanistan mein Aman ka qiyam nazdeek Hai taham malik se ghair Mulki foujon ke inkhala mein jald baazi danai nahi ho Gi .sanacher ko shaya honay waly is column mein Imran Khan ka kehna tha ke mazakraat ki maiz par aik pur Aman tawaquf Jung ke</td>
</tr>
</tbody>
</table>
4. Experiment

There are loads of utility That had been developed for fuzzy logic due to its ease and flexibility to deal with indistinct and uncertain facts. A fuzzy logic machine maybe counted because of the mapping of nonlinear enters information to scalar output facts.

Each part of the Fuzzy Logic is discussed below.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>i.</td>
<td>Fuzzification interface</td>
</tr>
<tr>
<td>ii.</td>
<td>Inference engine</td>
</tr>
<tr>
<td>iii.</td>
<td>Fuzzy ruled based</td>
</tr>
<tr>
<td>iv.</td>
<td>Defuzzification interface</td>
</tr>
</tbody>
</table>

4.1 Fuzzifier

In this section, the text capabilities are given as a difficult enter and convert that input into language values using a membership. The current approach uses the triangular membership feature for each feature and is separated into 3 fuzzy sets: less, medium, and greater.

4.2 Inference engine

In this section of the FLS take place the real process of plotting from given input to output the use of an association of IF-THEN rule outline in the guidelines base. First of all, the rule of thumb, the base is a defined sentence.

4.3 Fuzzy ruled based

Some Fuzzy rules are defined here.

IF (title word is more) and (thematic score is more) and (pronoun is average) and (numerical facts is more) and (TF-IDF is more) and (sentence length is medium) then (the sentence is very important).IF (title word is more) and (thematic score is more) and (pronoun is average) and (numerical facts is less) and (TF-IDF is more) and (sentence length is medium) then (the sentence is important).IF (title word is less) and (thematic score is less) and (pronoun is more) and (numerical facts is less) and (TF-IDF is less) and (sentence length is short) then (the sentence is unimportant).

4.4 Defuzzification

It generalized three-cornered Organization feature is used as an output membership characteristic as shown below in Equation 1.

$$C(x, y) = \left(\frac{a + b + c}{3}, \frac{l + m + n}{3}\right)$$

Whereas a, b and c are the usual values of the low, medium, and high respectively, and l, m, and n is the computed values of essential, common, useless, and very critical respectively. The structure of the Fuzzy Logic Model is as follow Figure 2.

Figure 2. Shows the flow chart of fuzzy logic.

4.5 Redundancy removal using Cosine similarity

The measurement of two texts documents or
articles resembling each other in terms of perspective and meaning, an important tool is known as Cosine is generally used. In Natural language Processing (NLP) an important metric is usually used to test whether there is any text similarity in the two records. The elimination of multiple sentences that contain similar documents from multiple document sources, the first step involves "the minimization of redundant information from the final review". Cosine similarity is one of the maximum used parallel measures in textual content summarization

Which is explained below.

The Cosine similarity measures the resemblance among pair of sentences $S_i = \{w_{i1}, w_{i2}, \ldots, w_{in}\}$ and $S_j = \{w_{j1}, w_{j2}, \ldots, w_{jm}\}$ is computed using Equation 2\(^{(32)}\).

$$\text{Sim} (S_i, S_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}}$$

(2)

In which $i$ and $j = 1, 2, 3, \ldots, n$ and $n$ are the whole numbers of sentences. Here, $w_{ik}$ and $w_{jk}$ is the burden of corresponding terms $t_k$ in sentences $S_i$ and $S_j$. It is the grouping of TF and ISF wherein TF is how frequently term $t$ appears in a sentence and ISF is what number of sentences of the summary group contain the period $t$. The weights have calculated the use of Equation 3.

$$w_{ik} = T F_{ik} \cdot \log n/nk$$

(3)

Where $TF_{ik}$ amounts to how many periods the terms $t_k$ appears inside the sentence $S_i$ and $n_k$ refer to the number of sentences containing the period $t_k$ in our experimentation, we have carried out the cosine similarity degree to the excessive recording sentences extracted by using the anterior phase. Then we have a look at the following sentence and examine it with a sentence (s) already include inside the summary sentences list. Those sentences that are too similar to the already protected sentences (Cosine similarity value is greater than threshold price 0.8) are considered redundant sentences and aren't to be involved within the summary. This manner is continued until the summary reaches the desired duration. We have used the python device to estimate the cosine similarity among two sentences (sci-learn,2019).

5. Evaluation Method

In this section, we evaluate our model result with the support of Rough Score and BLEU Score.

5.1 Rouge

Rouge (Recall Oriented Understudy for Gusting Evaluation) has come to be the standard technique that wants human-produced summaries which are observed as Gold standards and are related to the machine-created summaries. Rouge\(^{(32)}\) uses the n-gram statistics method to measure the Precision, Recall, and F-measure of a summarizer quantitatively. Below are the equations of these measures.

5.1.1 Precision

Precision is the ratio of retrieved record which might be associated with the question its miles calculated through calculating by way the usage of the following method.

$$\frac{|\text{relevant records} \cap \text{retrieved records}|}{|\text{retrieved records}|}$$

(4)

5.1.2 Recall

The recollection is the ratio of related which can be retrieved correctly. It is taken into consideration with the aid of the usage of the resulting method.

$$\frac{|\text{relevant records} \cap \text{retrieved records}|}{|\text{relevant records}|}$$

(5)

Precision and recall are inversely proportion:

As Precision increases $\uparrow$ recall $\downarrow$ decreases

Conversely

As Precision decreases $\downarrow$ recall $\uparrow$ increases

5.1.3 F_measure

$F$ degree is designed using calculating the accuracy and remember. It is a harmonic imply of accuracy and recall. We used the following formula for calculating the $F$ Score:

$$\frac{\text{precision} - \text{recall}}{\text{precision} + \text{recall}}$$

(6)

5.2 BLEU Score

BLEU (Bilingual evolution understudy) is one of
the fundamental estimation measures in the system translation area. It used an advanced form of precision-recall to match output text against many position sentences. The main programming in the BLEU implementer is to fit the n-grams of candidates with n-grams of position without seeing the region of the word. we have done our experiment through unigram.

In Equation 7, we show the scientific formula for BLEU Score.

\[ P = \frac{M_{\text{max}}}{W_t} \]  

(7)

For Example:

<table>
<thead>
<tr>
<th>Articles</th>
<th>Machine summary</th>
<th>Human summary</th>
</tr>
</thead>
</table>
| dengue, Alamat, ahthyat our ilaaj chhdsal qabala Pakistan mien taizi key sath muta-Arif hony wali wabai marz dengue ke mutaliq mukammal malomat sab ke liye inteihai zarore Hai taakay Kisi Bhi emergency Ki soorat mien khud kwavr –Apne pyaron nuqsan se Kisi Bhi museebat ke mutaliq paishgi maloomat ke liye inteihai hoten Hain. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai. bdalty Mausam ke sath hi dengue afzaish aur hamlon mien hojatahe is liye bar waqat dengue mutaliq aagahi hasil Hai hum isi mauzo par baat Karen ge baat Karen ge dengue Ki aqsam, tashkhees, ilaaj avrahtyat ke hawalay se avrpori koshish hogi ke qaryin tak mukammal maloomat pahonch jayen (dengue ki aqsam aur phelao) dengue virus ki family ke mumbar Hain aur un ka jeans virus Hai.

6. Result and Discussion

The advised Fuzzy Logic method is evaluated using the Rouge and BLEU Score method for the evaluation of 5 different articles. This Article is taken from News Articles dataset which belongs to different categories such as Sports articles, Poetry articles, Stories articles Education articles, and Health articles our fuzzy logic shows such a good result when we compare it with the human being-created summary. The red line show in a graph BLEU Scores while the BLEU color show precision, purple color display recall, and green color show f-measure. The BLEU Score and Rough Score give us good result against each summary. their evaluation results as shown below in Figure 3.
7. Conclusion

Text summarization has become a significant research domain as the information put away in the cloud is increasing gradually. The graph-based model, sentence-based model, the word-based model is some of the branches of extractive summarization. Based on Fuzzy logic single document summarization which routines a numerical feature-based model to contract with the incorrect and indecision of feature weight and generate a summary. During the whole experiment, we compare the system-created summary with other summarizer tools as mentioned which show us that our meant methods present better results than other summarizers in the period of BLEU Score, Precision, Recall, and F-Measure Method.

References


