Digital Object Identifier

Text summarizing of Urdu using Extraction Base Technique

MUHAMMAD UZAIR[1], HIRA SALEEM[2],QAZI DANISH AYUB.[3] , and AHMAD NAWAZ[4]

ABSTRACT Pre-processing is preliminary stage in several fields counting IR and natural language processing. The outcome of basic pre-processing settings on English for matter content summarization is well studied. However, there is no such effort found for the Urdu language (with the quality of knowledge). In this study, we investigate the effect of basic pre-processing settings for single-document text summarizing for Urdu, on a target corpus using several trials. The analysis is performed using the state of the art algorithms for extractive summarizing and the effect of stop word removal, lemmatizing, and stemming is analyzed. Results showed that these pre-processing settings develop the effects.

I. INTRODUCTION

Urdu language is member of an Indo Aryan group which is widely spoken. Urdu script is an extended version of Perso-Arabic script. Similar to the other Perso-Arabic scripts, it is written in RTL(right-to-left) system, in a complex, cursive-style writing systems. Urdu become heir to Arabic Persian and the inherent languages of South Asia using a lot of their vocabulary. Due to the impact of this influence, Urdu has a complex structure. In terms of grammar, it has a comparatively loose word order (Subject Object Verb). Regardless of spoken by millions of individuals, Urdu is an under-resourced language in terms of accessible computational resources.

The accessibility of benchmark corpora plays a significant part in the advancement of apparatuses and methods for different NLP undertakings. For a synopsis of programmed textual content, shared assignments offered by report understanding The conference (DUC) and the text analysis convention (TAC) provide numerous suitable-nice reference corpus arrangements generally for English.These corpora comprise of single and multi-record summaries composed by people. These outlines are abstractive just as extractive. These arrangements of benchmark corpora have been utilized for the evaluation and assessment of summarization frameworks (chiefly for English) and results are distributed routinely.

Inaccessibility of standard assessment assets is one of the fundamental obstructs for doing research in the Urdu language. As an initial step, this work builds up a benchmark Urdu summary corpus. It contains 50 articles and their relating abstractive summaries covering different areas. These are human-composed single document abstracts. There is just a single summary for an article.

For the most part, programmed text summarization tends to the issue of producing short data text a solitary document (or numerous document composed on a similar topic). Generally, this shortened content is essentially not exactly the document text however, never greater than half of the source text. Two main techniques used for automatic text summarization are abstraction based and extraction based. Abstraction based approach separates key focuses introduced in various sentences. furthermore, develops a cognizant summary from the source text by disposing of unimportant subtitles. This requires tackling hard inquiries, for example, semantic portrayal, surmising, natural language generation, and so on.

Interestingly, extraction based strategy outline moderately straight forward information from text. It depends on distinguishing most significant sentences from the source document and assign them weights (in light of their significance). Summary is then formed utilizing top n sentences utilizing these weights. The estimation of n relies upon the length of required summary. These chosen sentences are kept unblemished as units even when a portion of their parts may not contain most significant data.

II. LITERATURE REVIEW

Automatic text summarization tool facilitate the user to understand the whole document in a short time of period. Computerized textual content summarization is a critical tool on this age Statistics overload may be used to:

Summaries within the each day lifestyles cycle.

- Titles inclusive of summaries of newspaper guides
- Table of contents, which include a precise of a e-book, magazine

- Digest: precise of stories
- Highlights - precise of a meeting
- Summary of this type of abstract of a systematic article
- Ads along with weather forecast, stock market, News
- Biography together with curriculum vitae, obituary
- Synopsis of books
- Condensed assessment of books, music, performs, and so on.
- Descending scales like maps, thumbnails

We are living in a global village, there are many languages spoken around the world there are two ways of communication: first, in person, and second, in written. In writing communication, translators have to translate documents in another language. Instead of full document translation, a translation summary is usually sufficient. Therefore, an automatic summarization tool may facilitate the translator. With the increase of World Wide Web, a huge amount of data is described in web pages. If it also shows the summarized data then it will be helpful for the reader to understand the whole text in a short form which is helpful for smart phones graphical visualization display memory problem.

For extraction based strategy, a few unsupervised techniques have been proposed throughout the long term. Some exemplary models are (Luhn, 1958) and (Baxendale, 1958; Edmundson, 1969) that utilization word frequency and sentence position individually. TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004) are among the renowned and broadly utilized chart based algorithms(Leite et al., 2007). We have utilized the two of them in this work for the assessment of pre-processing settings. To the most amazing aspect our insight, there are only few studies which investigates the impact of pre-processing settings. Summary extraction algorithm are by and large language independent. In any case, customized pre-processing work, which is language dependent, is constantly required particularly for morphologically rich dialects. Instances of such pre-processing work could be distinguishing legitimate word limit and sentence limit. Likewise, language-subordinate resources, for example, stemmers, the arrangements of stopwords, lemmatizers, and so on, may likewise be needed to improve the nature of a generated summaries. In this manner, examining pre-processing settings for Urdu, which is for sure morphologically rich, is a significant research topic that this paper attempts to explore.

III. DATA COLLECTION AND METHODS

Urdu Summary Corpus (Humayoun et al., 2016) is utilized in this study which is a benchmark corpus for single document abstraction. It gives 50 articles and their comparing human-composed summaries (abstracts) covering different domains. All the more decisively, Urdu Summary Corpus comprises of:

1) fifty Urdu articles that were gathered from different sources, and normalized

2) fifty abstractive single document outline

(3) fifty grammatical feature labeled articles

4) fifty morphologically dissected articles

(5) fifty lemmatized articles

(6) fifty stemmed articles.

The algorithms utilized in this examination are

1. Latent semantic analysis(LSA)
2. TextRank
3. LexRank
4. SumBasic
5. Luhn

The initial three are proposed by (Erkan and Radev, 2004) and the fourth calculation is proposed by (Mihalcea and Tarau, 2004). As a pattern, Leadbased strategy is utilized in which basically first n sentences are chosen from the document as a summary.

Following is the preprocessing settings implemented on documents in the papers. There are pre-prepared three lists of stopwords for the experiments.

- The principal list is taken from (Burney et al., 2012). It contains 519 words.
- The subsequent documents is worked by computing the term frequecy (Kenney and Keeping, 1962; Lo et al., 2005), on an enormous Urdu corpus (Jawaid et al., 2014). It contains 500 words.
- Some examinations stress the utilization of redid stopwords lists that ought to be extricated from the domain corpus of the undertaking close by (Lo et al., 2005; Blanchard, 2007). It likewise contains 500 words.
- The fourth rundown contains just closed classes and has 195 words. The list is produced from the open source assets of Urdu morphology (Humayoun et al., 2007). We are utilizing the open-source assets of Urdu morphology (Humayoun et al., 2007)

According to paper by apply lemmatization and stemming, results does not improve on urdu language so we generate our results on different pre-processing approaches. Results of pre-processing is shown in Firgure 1.

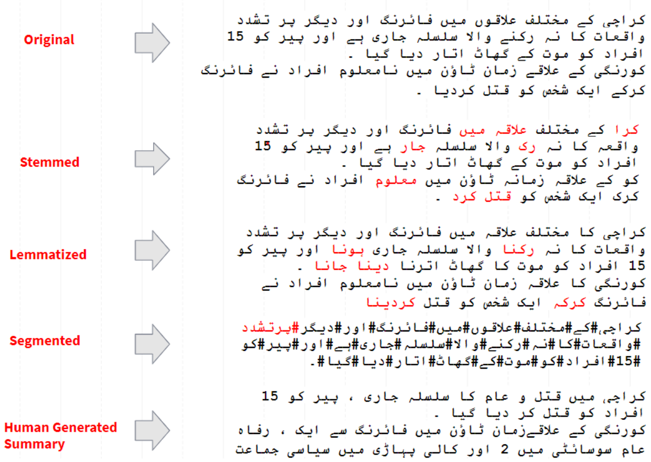
- Unsegmented (Not Proper Tokenize)
- Unsegmented (Not Proper Tokenize)+Stemming
- Segmented (Proper Tokenize)
- Segmented (Proper Tokenize) +Stemming

Word segmentization is used for abstraction in this study. A stemmer creates a shortened structure for all arched surface types of a word.

IV. EVALUATION AND RESULTS

Following evaluation algorithms are used in this study:

- Precision
- Recall
- F1score
- Cosine Similarity
- Unit Overlap
- Rouge1
- Rouge2
- RougeL Sentence Level
- RougeL Summary Level



کراچی کے مختلف علاقوں میں فائرنگ اور دیگر پر تشدد
واقعات کا نہ رکنے والا سلسلہ جاری ہے اور پیر کو 15
افراد کو موت کے گھاٹ اتار دیا گیا .
کورنگی کے علاقے زمان ٹاؤن میں نامعلوم افراد نے فائرنگ
کرکے ایک شخص کو قتل کردیا .

کرا کے مختلف علاقہ میں فائرنگ اور دیگر پر تشدد
واقعہ کا نہ رک والا سلسلہ جار ہے اور پیر کو 15
افراد کو موت کے گھاٹ اتار دیا گیا .
کو کے علاقہ زمانہ ٹاؤن میں معلوم افراد نے فائرنگ
کرک ایک شخص کو قتل کرد .

کراچی کا مختلف علاقہ میں فائرنگ اور دیگر پر تشدد
واقعات کا نہ رکنا والا سلسلہ جاری ہونا اور پیر کو
15 افراد کو موت کا گھاٹ اترنا دینا جانا .
کورنگی کا علاقہ زمان ٹاؤن میں نامعلوم افراد نے
فائرنگ کرکہ ایک شخص کو قتل کردینا

کراچی#کے#مختلف#علاقوں#میں#فائرنگ#اور#دیگر#پرتشدد
#واقعات#کا#نہ#رکنے#والا#سلسلہ#جاری#ہے#اور#پیر#کو
#15#افراد#کو#موت#کے#گھاٹ#اتار#دیا#گیا#.

کراچی میں قتل و عام کا سلسلہ جاری ، پیر کو 15
افراد کو قتل کر دیا گیا .
کورنگی کے علاقےزمان ٹاؤن میں فائرنگ سے ایک ، رفاہ
عام سوسائٹی میں 2 اور کالی پہاڑی میں سیاسی جماعت

FIGURE 1. Pre-Procesing

ROUGE-1 refers to overlapping unigrams (every word) between the gadget summaries and the reference. ROUGE-2 refers to the overlap of bigrams among abstracts machine and reference. ROUGE-Lis Statistics Focused on Longest Common Subsequence (LCS). Sentence level structure similarity is naturally taken into account by the longest common subsequence problem and immediately defines the longest co-occurrence in sequence n-grams. Retirement is determined via | gadget-human options overlay | / | selected terms in step with human |. And precision is the fraction of adequate formula frame sentences: | system-human preference overlay | / | terms chosen through the device |. The F1 Score is the Precision and Recall Weighted Average. Cosine Similarity is commonly used in text analysis to calculate document similarity.

This paper reports Extraction based analysis for Urdu text summarization, using multiple state of the art algorithms and different pre-processing techniques. Extraction base summary provides some specific lines from the whole text which represent the summary of the whole document in short lines. We have freely available benchmark fifty articles of Urdu summary corpus were collected from various online sources mainly news portals and blogs in form of (1) Processed (2) Processed-Seg with stemmed and without the stem (Humayoun et al., 2016). This may facilitate the researcher to check the impact of processed un-segment (words separated with space) and processed-seg (words separated with specific symbol) in the language processing of Urdu. We used five state art of algorithms (Luhn,TextRank,LexRank,SumBasic and LSA) in these analyses which provide marginally different results. However, we take the average of almost Nine evaluation measures (Rouge 1, 2, L, Recall, Precision, F1, Unit Overlapping and Cosine Similarity) in which processed un-segment text with stem and without stem have same results but also have properly processed-segment with stem and without stem same results but varies from each other. We observe similar as mentioned in the paper a small study of Urdu Summary Corpus (Humayoun et al., 2016) with or without stemming have no changes in results but processed un-seg and processed-seg produce better results on the textual summary as compared to stemming. However stemming has no impact on Urdu textual summary, but proper processed-seg and un-seg have the most significance in text summarization and can also check evaluation in the given table. After analysis of all algorithms and evaluation measures we able to seen Luhn (1) TextRank(2) algorithm have good results as compared to other algorithms and Cosine Similarity, Rouge 1,2 ,L Unit Overlapping are good evaluation measures as compare to F1,Recall and Precision. Table1 and Table 2 shows the average of all evaluation measures of the LSA and TextRank algorithm.

Evaluation Measures	Processed	Processed + Stem	Processed_Seg	Processed_Seg + Stem
F1	0.001367	0.001367	0.000000	0.000000
Recall	0.001367	0.001367	0.000000	0.000000
Rouge 1	0.491816	0.491816	0.527816	0.517816
Rouge 2	0.260490	0.260490	0.280980	0.278959
Precision	0.001367	0.001367	0.000000	0.000000
Unit Overlap	0.296143	0.296143	0.305592	0.305592
Rouge Sent. L	0.334633	0.334633	0.338000	0.338000
Rouge Sum. L	0.011245	0.011245	0.011245	0.011245
Cosine Similarity	0.694980	0.694980	0.788551	0.788551

TABLE 1. Conclusion Table of Evaluation Measures with Different Data Processing (LSA Algorithm)

Evaluation Measures	Processed	Processed + Stem	Processed_Seg	Processed_Seg + Stem
F1	0.001143	0.001143	0.000000	0.000000
Recall	0.001143	0.001143	0.000000	0.000000
Rouge 1	0.523714	0.523714	0.510551	0.510551
Rouge 2	0.290878	0.290878	0.278959	0.278959
Precision	0.001143	0.001143	0.000000	0.000000
Unit Overlap	0.304490	0.304490	0.301980	0.301980
Rouge Sent. L	0.329429	0.329429	0.332796	0.332796
Rouge Sum. L	0.014224	0.014224	0.010551	0.010551
Cosine Similarity	0.712469	0.712469	0.792224	0.792224

TABLE 2. Conclusion Table of Evaluation Measures with Different Data Processing (TextRank Algorithm)

V. CONCLUSION

With or Without Stemming Results does not change (We observe similar as mention in paper).Proper Segmentation Produce better results on urdu data(We observe similar as mention in paper). Luhn(1) TextRank(2) Algorithm have Good Results as compared to other's.Cosine Similarity,Rouge 1,2 ,L Unit OverLapping are Good Evaluation measures as Compare to F1,Recall and Precision Unit Overlapping = Rouge L Sentence Level + Rouge L Summary Level

REFERENCES

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958, doi: 10.1147/rd.22.0159.

[2] Humayoun, Muhammad, and Hwanjo Yu. 'Analyzing Pre-Processing Settings for Urdu Single-Document Extractive Summarization'. Proceedings of the Tenth International Conference on Language Resources and Eval-

uation (LREC'16), European Language Resources Association (ELRA), 2016, pp. 3686–93. ACLWeb, https://www.aclweb.org/anthology/L16-1585.

[3] Humayoun, Muhammad, et al. 'Urdu Summary Corpus'. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), 2016, pp. 796–800. ACLWeb, https://www.aclweb.org/anthology/L16-1128.

[4] Belica, Mišo. Sumy: Module for Automatic Summarization of Text Documents and HTML Pages. PyPI, https://github.com/miso-belica/sumy. Accessed 6 Feb. 2021.

[5] Humayoun, Muhammad. Humsha/USCorpus. 2016. 2020. GitHub, https://github.com/humsha/USCorpus.