Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.

ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-67

Outlier Document Filtering Applied to the Extractive Summarization

[Metin Turan , Coşkun Sönmez]

Abstract— **Summarization requires selection of the more informative sentences within a set of documents. Generally, process assumes the document set includes related topics to a subject. However, some of the documents may be outlier and the effect of an outlier document might affect the success of extractive summary.**

Research is focused on filtering documents at the extraction stage these are outlier. Extraction finds the outlier documents far distance from representative document set word vector (DSWV).

DUC 2006 data set is used for tests. System summaries are compared with the systems generated by DUC participants. Results points out that filtering outlier documents overwhelm all the systems fairly.

Keywords— Document Processing, Extractive Summarization, Outlier Detection, Similarity Measure

I. Introduction

Document processing age is still not capable of extracting information as a human reader. Moreover, the importance of content in the document may also vary from one reader to another. Automatize the process needs to use both textual properties and resolve the language structure which is a really colicky operation with nowadays techniques.

In other words, text summarization can be either extractive or abstractive [1]. In extractive summarization, important sentences are identified and extracted directly. This type of summarization is the interest area of this article. Abstractive summary requires linguistic support which is not a part of this work.

The textual properties are used to construct the representative vector for a unit (sentence, paragraph, section, document) in text. The best descriptive property found in the text is a simple word.

The Luhn's [2] work is accepted a milestone for document summarization. It is the first work using term frequency (TF).

MSc. Metin TURAN
Yıldız Technical University
Turkey

Prof. Dr. Coşkun SÖNMEZ
İstanbul Technical University
Turkey

After one decade later Edmundson [3] suggested three methods (cue, title, location) to support frequency additionally when evaluating the sentence weights. Normalization for all units in text (how often it appears in units) is also realized and achieved by using Inverse Document Frequency (TFIDF) measure instead of TF. Multi document summarization approaches worked in the survey published by Kumar [4].

Similarity or distance of units[5] is another research area to obtain extractive summary. Researchers tend to cluster similar units [6, 7, 8, 9] (generally sentences) to prevent overlap or to find important units.

Common similarity measures are researched for effectiveness in Anna's work [10]. Term Weighting functions and a learning model for similarity is also worked by Wen-tau Yih [11]. Also new similarity measures are proposed by some researchers [12, 13, 14].

Although extractive summary is obtained from related set of documents, it is not guaranty that a few documents can be out of general subject or they may be considered as outlier. Outlier detection is generally applied for cleaning data set or clustering data [15].

Yu Nie and colleagues [16] work sentence filtering to maximize a global search criterion. It is not a completely an outlier application, however the idea is to remove most irrelevant sentences from sentences set could be considered as finding outlier sentences.

Find outlier documents and eliminate them for summarization has not been considered as a technique in the literature yet. In this work an algorithm for outlier document filtering and a similarity metric for sentence ordering is suggested. A system is developed and results are evaluated.

The article is organized so that, the next chapter summarizes the inspired and related work. The chapter III includes data and evaluation technique. The chapter IV gives experiment results in comparison with the systems results produced in the DUC 2006 competition. The last two chapters discuss the results and further works respectively.

II. Work

The idea behind the finding outlier documents is coming from reality of unbalanced document set. Whenever a document is selected for a subject, it is generally evaluated by its header, sub topics however it is not investigated for details covered. Moreover, writer chooses the content of a document,

Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.

ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-67

so details depend on the writer ability, focus and concern.

Human summarizer focuses on the relevant data, so that irrelevant data is non-sense. He/she does that with intuition and heuristically. However, if we don't concern with this issue, system output would be affected by noisy information.

The information for outlier system is the words and their frequencies obtained from documents. So, we focus on the word dispersion through documents. If a word would be seen at least a predetermined ratio of documents, it would be considered as a representative word for all document set. Otherwise it is accepted as noisy information.

Meaningful words [17] and word dispersion [18] are considered in the literature. Goodman [19] also developed a procedure to determine informative words by frequency and position. However, they are only limited to the word selection. Document filtering has not been considered yet.

Experiments are divided into three separate word dispersion ratio (WDR) as %25, %50 and %75. For example if a document set (DS) contains 25 documents and WDR is %25 , then a word must be referred in at least $\lceil 25 * \%25 \rceil = 7$ documents to be selected as document set word vector(DSWV).

DS = { d_1 , d_2 , … , d_{max} }

The words provide the WDR rule is used to construct DSWV. Distances between documents word vectors (DWV) and DSWV are calculated. The documents these are far away from the 2*σ distance on both direction from average distance (μ) are marked as outlier.

Outlier system is developed by two stages. First stage is called processing and second stage is called extraction.

Processing stage parses each document separately. Each word is preprocessed for eliminating the stop words and finding the stems (Porter stemming is used). Sentences are determined. DWV and sentence word vectors (SWV) is constructed for document and sentences in the document.

DWV is a word list contains the words in the document after stop words elimination. SWV is also a word list contains the words in the sentence after stop word elimination. Where, DSWV includes all different existing words in the DWV's.

DWV_i = { w_i , w_j , w_k , w_m , w_y }, SWV_{i1} ={w_k , w_y}

Extraction stage is composed of two phases. Former applies the the pseudo code given algorithm called MarkOutlier() (Algorithm 1) to find outlier documents and constructing the DSWV during the extraction phase. The parameters, documentNumber and dispersionRatio, are used for the total document count in the DS and WDR selection respectively.

Latter finds the similarity of each SWV with the DSWV. The documents which are marked as outlier wouldn't be considered for further processing. The further similar sentences are selected for extractive summary under restriction of the summary size.

MarkOutliers(float dispersionRatio, integer documentNumber)
{
#minDocument =$\lceil$ dispersionRatio * documentNumber$\rceil$

/* construct document set vector */
for each word w_i in DSWV
search in { DWV_1 , DWV_2,…, $DWV_{documentNumber}$}
if (w_i is a member of at least #minDocument document)
Set w_i ϵ DSWV;

/* calculate distance of each document */
for each d_i in the DS
Calculate euclidean distance (ED_i) of DWV_i with DSWV;

/* find outlier boundary using μ and 2*σ */

distance_μ= $\frac{\sum_{i=1}^{documentNumber} ED_i}{documentNumber}$;

limit_μ_+_2σ = distance_μ + 2*σ;

limit_ μ_-_2σ = distance_μ – 2* σ;

/* mark outlier documents */
for each d_i in the DS
if (ED_i < limit_ μ_-_2σ OR ED_i > limit_μ_+_2σ)
mark d_i as outlier;
}

Algorithm 1 MarkOutlier Algorithm

The following similarity function, we called **match percent** (1), is used to calculate similarity of each SWV with the DSWV.

$$\textbf{match percent} = \frac{count\ of\ true\ words}{count\ of\ words\ in\ sentence\ word\ vector} \quad (1)$$

Count of true words is the number of words in the SWV which matches with the DSWV.

Match percent(1) is structured to eliminate the superiority of long sentences (word count is much more than a shorter one) over short sentences. Match percent is a ratio of similarity success of sentence words.

III. Data and Evaluation

The DUC 2006 corpus includes 50 document set from Financial Times of London and Los Angeles Times. Each document set includes 25 news. The system summary is limited to 250 words for each document set. Four models are used to produce human summaries for each document set. Thirty-five systems were attended to the competition and all

Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.

ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-67

generated an output for each document data set. They are evaluated, and scores are published.

ROUGE metrics [20] are used to evaluate each system generation with human models. ROUGE metrics included in DUC 2006 tests as follow:

ROUGE-N: N-gram based co-occurance statistic is given. DUC tests are done from 1 to 4 gram.

ROUGE-L: Longest common subsequence (LCS) based statistic is given. It is sentence based similarity and identifies longest n-gram sequence.

ROUGE-W: consecutive LCS based statistic is given.

ROUGE-SU: Bigram plus unigram skipped co-occurance statistic is given.

ROUGE metrics include three properties: precision(2), recall(3), F-Score(4). Their definitions are given as follows:

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (2)$$

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (3)$$

$$\text{F-Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

IV. Experiment

Outlier system runs for every WDR on each data set separately. The average ROUGE score is calculated for 4 models. In other words, one data set requires 3 runs, totally (all data set) 150 runs.

Table-I gives the average of participants systems and outlier system metrics obtained from ROUGE applying all data sets in DUC 2006.

V. Conclusion

Every system processing data requires clean data, in other words removing noisy data, to get a better result. We couldn't decide which document is outlier until all documents are processed. It is pointing out that why we filter the documents these are outlier at the extraction stage.

In this work, applying well known technique- outlier detection- is suggested for extractive summarization. Moreover a new linear similarity measure (match percent) is proposed and applied.

The obtained results show that finding and eliminating outlier documents from summarization helps to get more similar system outputs to the human summaries.

At the ROUGE-1 level it seems that outlier system nearly doubly successful on comparison with DUC participants systems. The separation between systems can be seen at the next ROUGE levels (n-grams). The ROUGE scores drop down dramatically for DUC participant systems for 2 to 4 grams. However, outlier system outputs for n-grams don't drop sharply. Moreover, it is really interesting that 4-gram scores for outlier system are nearly equal to the DUC participant systems 1-gram scores. This is fairly a good benchmark.

TABLE I. Comparison of DUC 2006 participants and outlier system ROUGE scores for different *WDR*'s

	DUC-participants	Outlier-*WDR* 25%	Outlier-*WDR* 50%	Outlier-W*DR* 75%
ROUGE-1				
Average-R	0.371	**0.641**	0.635	0.637
Average-P	0.386	**0.540**	**0.540**	0.535
Average-F	0.377	**0.583**	0.581	0.579
ROUGE-2				
Average-R	0.073	**0.413**	**0.413**	0.417
Average-P	0.076	0.347	**0.351**	0.349
Average-F	0.074	0.375	0.377	**0.379**
ROUGE-3				
Average-R	0.020	0.344	0.344	**0.348**
Average-P	0.021	0.289	**0.292**	0.291
Average-F	0.021	0.312	0.315	**0.316**
ROUGE-4				
Average-R	0.008	0.301	0.302	**0.305**
Average-P	0.009	0.253	**0.256**	0.255
Average-F	0.008	0.273	0.276	**0.277**
ROUGE-L				
Average-R	0.340	0.568	0.571	**0.574**
Average-P	0.353	0.478	**0.485**	0.481
Average-F	0.346	0.516	**0.522**	**0.522**
ROUGE-W				
Average-R	0.099	0.156	0.157	**0.158**
Average-P	0.188	0.256	**0.260**	0.258
Average-F	0.129	0.193	**0.195**	**0.195**
ROUGE-SU4				
Average-R	0.128	**0.408**	0.399	0.401
Average-P	0.133	**0.293**	**0.293**	0.285
Average-F	0.130	0.334	0.332	0.330

Proc. of the Intl. Conf. on Advances In Information Processing And Communication Technology - IPCT 2014.

ISBN: 978-1-63248-021-7 doi: 10.15224/ 978-1-63248-021-7-67

On the other hand, WDR values effects the ROUGE scores slightly. They are nearly same for %25, %50 and %75 WDR's. This shows that if WDR (between %25-%75) is applied to a summarization system, extractive summary would produce nearly same results.

VI. Further Work

Firstly, **2*σ** distance on both side covers nearly 95 percent of all data for normally distributed data set. In other words, we only filter %5 of the documents by using outlier filtering. Using **1.5*σ** distance instead of **2*σ** distance in the MarkOutliers() algorithm could be suggested. It would filter nearly %13 of the document set. It may produce considerable better result on the unbalanced data intuitively.

Secondly, instead of suggested similarity technique- match percent- any other similarity measures could be applied and evaluated.

Additionally and more importantly, outlier filtering technique could also be applied instead of match percent similarity function. By this way, some irrelevant sentences could also be eliminated at the extraction stage. Possibly, the order of sentences would be more consistent.

References

[1] Radev, D.R., E. Hovy, K. McKeown, "Introduction to the Special Issue on Summarization" , J. Comput. Linguistics, 2002, pp. 399-408.
[2] Hans Peter Luhn, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, 1958, pp. 159-165.
[3] H. P. Edmundson, "New methods in automatic extracting", Journal of the ACM, 1969, p.p. 264-285.
[4] Yogan Jaya Kumar, Naomie Salim, "Automatic Multi Document Summarization Approaches", Journal of Computer Science, 2012, pp. 133-140.
[5] Sung-Hyuk Cha, "Comprehensive Survey on Distance/ Similarity Measures between Probability Density Functions", International Journal of Mathematical Models and Methods in Applied Sciences, 2007, Issue 4 , Vol. 1, pp. 300-307.
[6] De-Xi Liu, Yan-Xiang He, Dong-Hong Ji, Hua Yang, "A Novel Chineese Multi-Document Summarization Using Clustering Based Sentence Extraction", Proceedings of the 5th International Conference on Machine Learning and Cybernetics, 2006, pp. 2592-2597.
[7] Maruthamuthu, Maheedharan, Kirubakaran and Shanmugasundaram Hariharan, "Experiments on Clustering and Multi-Document Summarization", The Icfai University Press, 2009,pp. 64-73.
[8] Ramiz M. AliGuliyev, "Clustering Techniques and Discrete Particle Swarm Optimization Algorithm For Multi-Document Summarization", Computational Intelligence, 2010,vol.26. Num.4, pp. 420-449.
[9] Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, Yihong Gong,"Integrating Document Clustering and Multidocument Summarization", ACM Transactions on Knowledge Discovery from Data, 2011, Vol. 5, No. 3, Article 14.
[10] Anna Huang, "Similarity Measures for Text Document Clustering", New Zealand Computer Science Research Student Conference, 2008, pp. 49-56.
[11] Wen-tau Yih, "Learning Term-weighting Functions for Similarity Measures", Conference on Empirical Methods in Natural Language Processing, 2009, pp. 793-802.
[12] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, "Quick Asymmetric Text Similarity Measures", Proceedings of the Second International Conference on Machine Learning and Cybernetics, 2003, pp. 374-379.
[13] Wen-tau Yih, Christopher Meek, "Improving Similarity Measures for Short Segments of Text", Association for the Advancement of Artificial Intelligence, 2007, pp. 1489-1494.
[14] Ramiz M., Aliguliyev,, "A new Sentence Measure and Sentence Based Extractive Technique for Automatic Text Summarization", Expert Systems with Applications, 2009, pp. 7764-7772.
[15] Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook Chapter I, Kluwer Academic Publishers, 2005, pp. 1-16.
[16] Yu Nie, Donghong Ji, Lingpeng Yang, Zhengyu Niu, Tingting He, "Multi-document Summarization Using a Clustering-Based Hybrid Strategy", AIRS , 2006, pp. 608-614.
[17] Helen Balinsky, Alexander Balinksky, Steven Simske, "Document Sentences as a Small World", Systems Man and Cybernetics(SMC), 2011, pp. 2583-2588.
[18] Meng Wang, Xiaorong Wang, Chungui Li, Zengfang Zhang, "Multi-Document Summarization Based on Word Feature Mining", International Conference on Computer Science and Software Engineering, 2008, pp. 743-746.
[19] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, Hisami Suzuki, "Multi-Document Summarization by Maximizing Informative Content Words", IJCAI, 2007, pp. 1776-1782.
[20] Chin-Yew Lin,"RUGE:A Package for Automatic Evaluation of Summaries", In Proceedings of the Workshop on Text Summarization Branches Out(WAS), 2004.

About Author (s):

Metin TURAN
- Fifteen years experience in analyzing, designing, programming and project management.
- Nine years experience as lecturer at Computer Engineering Department of İstanbul Kültür University.
- He is currently a doctorate student and working as a consultant for big data problems.

Coşkun SONMEZ

Professor of Computer Engineering Department at Istanbul Technical University, Istanbul. He received the PhD degree at Cambridge University (UK). His main research interests are: Artificial Intelligence, Computational Intelligence, Robotics, Real Time Systems and Embedded System Applications.