Summarization of Table Citations from Text

Monalisa Dey[1], Salma Mandi[2], Dipankar Das[3],
[123]Jadavpur University, Kolkata
[1]monalisa.dey.21@gmail.com, [2]salmamandimandi@gmail.com,
[3]dipankar.dipnil2005@gmail.com,

Abstract

This paper reports our attempt to design a corpus for table content summarization, abstractive and extractive, where a table is cited in a scientific document. We have utilized 200 scientific publications in the computer science field covering 10 different domains like machine learning, automatic summarization etc. to construct the corpus. The dataset preparation for this work has been extremely daunting due to the nature of the data. The prepared dataset has been used for training, testing and evaluation. Manual annotators have been employed to validate the gold standard data in corpus. Moreover, we have also proposed two systems based on TF-IDF approach and Transition point approach to generate an extractive summarization system. The similarity score between the system generated summaries and gold standard data is calculated using standard metrics to evaluate the quality of the generated extractive summary. Finally, we have documented our observation have presented an error analysis of the system using standard metrics viz. BLEU and ROGUE.

1 Introduction

Authors use various non-textual components to represent information in a document or article. The most commonly used entities are tables to present findings or experimental results, graphical forms and figures for describing a process or presenting the output, flowcharts to depict the system flow, etc. These elements are a source of vital information and hence the importance of retrieving information from these components are increasing rapidly over the years. Moreover, often the most important experimental results and ideas in any article are presented using a table. A lot of time and effort can be saved if a researcher can understand the content of a table, without having to read the entire paper. It may also allow the researcher to examine more results that he would normally do. Consequently summarization systems play a major role in helping the reader extract critical information automatically and intelligently.

Most recently a few noteworthy contributions have been made in this area. ScienceDirect [1] is offering a table/figure preview feature in some of its articles. CiteseerX[2] is providing intelligent information extraction like figures, citations, pseudocodes etc. Consequently, summarization system becomes important to ease people in extracting the information automatically and quickly.

Although, table content summarization systems have many potential uses like summarizing a patient information from a table of symptoms and potential diseases, weather prediction form a table of daily weather reports, wikipedia infobox summarization, analysis of games like cricket,from their score tables etc., previous researches show that the challenge is finding a suitable corpus that can be used for training, testing and evaluating a table summarization system.

In order to address this issue, we have been motivated to start our study by constructing a corpus of table-content summarization in two forms, namely, extractive and abstractive summary using NLP based techniques. The techniques are, data preparation, information extraction, module building and validations. Extractive summary is obtained by extracting sentences from the article that describe the table, and abstractive summary is ob-

[1] www.sciencedirect.com/

[2] citeseerx.ist.psu.edu/

Proc. of ICON-2018, Patiala, India. December 2018 ©2018 NLPAI, pages 35–42

tained by extracting the captions associated with each table.

To construct the corpora, we have considered the following challenges and adopted feasible solutions;

A. ***How to obtain and pre-process the dataset?*** To address this challenge, we have downloaded 499 different tables from 200 computer science publications which covers various domains like Named Entity Recognition, Machine Translation, Machine Learning etc. Most of the articles are available in PDF format and hence we had to convert them to text format using for further processing purposes using PDFTextStream[3].

B. ***How to extract the table caption as abstractive summary?*** Extracting a caption is a challenge as it is written in various formats throughout different domains and writing styles. To address this issue, we have observed that a caption sentence for a table consists of FOUR parts. They are <*TABLE*>, <*INTEGER*>, <*DELIMETER*> and <*TEXT*>. Thus, to distinguish between a caption and the rest of the sentences, we propose that any sentence following the above-mentioned pattern is a caption, and the caption can form the abstractive summary of a table.

C. ***How to extract the reference text as extractive summary?*** Although captions provide details about the information in a table, it is quite possible that they might not contain enough information to assist a reader to interpret the content fully. To address this issue, we have extracted the text which is referencing the table within the document. In order to do so we have followed the same method as mentioned in the previous challenge with a few differences in the pattern. We have observed that a sentence in the vicinity of the reference sentence may provide accurate information about the context in which a table is used. Hence, we have also extracted and captured such contextually crucial sentences.

D. ***How to validate the obtained summarized output?*** The evaluation process of the system has been divided into two parts, namely, accuracy of summary identification and quality evaluation. For validating the first part, we have taken the assistance of two annotators namely, a manual annotator, A1 and our system. The Cohen's Kappa agreement analysis technique is then used to study the inter annotator agreement scores. For evaluating the quality of the summaries generated, we have employed a sentiment based similarity technique which generated a similarity score between the system generated and reference summaries which are again identified by A1. Moreover, the prepared corpus can be considered as a gold standard dataset. This is so because, to design the corpus, we are using captions and texts from the scientific publication written by the author himself.

E. ***How to present the output as a structured corpus*** To address this challenge, we have prepared an annotated corpus which contains information about various tables and their related features like abstractive summary, extractive summary, no. of rows, no. of columns etc.

The contributions of the task is to address the above-mentioned challenges, and present an annotated structured corpus with summarized output as a standard dataset for table content summarization.

The overall structure of the paper is as follows. Section 2 presents the related work carried out in this domain. Section 3 and Section 4 describe the dataset annotation in details and the model building. Section 5, Section 6 and Section 7 describe the evaluation process. Finally Section 8 describes the concluding remarks and future scope of the research.

2 Related Work

Table construction methods in free text are simple but the expressive capability is limited. The markup languages like HTML provide very flexible constructs for writers to design tables. The flexibility also shows that table extraction in HTML is harder than that in plain text. The task of table extraction from text document in (Ng et al., 1999) was recognizing table boundary,column and row. These are defined as three separate classification problem and relies on sample training texts in which the table boundaries, columns and rows have been correctly identified by human annotator.Machine learning algorithms are used to build classifiers from the training examples, one classifier per subproblem. This system is flexible and easily adoptable to text in different domain with different table characteristics. In (Wang and Hu, 2002) machine learning based approach has used for classification of table in HTML document as either genuine or non-genuine table. A set of novel features has defined which reflect

[3] 7http://snowtide.com/PDFTextStream

the layout as well as content characteristics of tables. For the table detection task , the decision tree classifier is used as here features are highly non-homogeneous.They also experimented with support vector machine which shows the best performance in text categorization.However, this system misclassified a table due to the ambiguous content e.g,a table contain many hyper-links which is unusual for genuine table.This is case where layout features and the kind of shallow content features are not enough. Deeper semantic analysis would be needed in order to identify the lack of logical coherence.

An automated Table Extraction approach used in (Tengli et al., 2004) that exploits formatting cues in semi-structured HTML tables, learns lexical variants from training examples and uses vector space model to deal with non-exact matches among labels.In (Chen et al., 2000) tables are mined from large scale HTML texts. This task composed of five modules:hypertext processing,table filtering,table recognition,table interpretation and presentation of results. Table filtering module filters out impossible cases by heuristic rules.Table recognition module recognize table by the content of the cells.Table Interpretation module interpret the table attribute-value relationship either column wise or row wise.Presentation of results module results the table in a sequence of attribute-value pairs.

3 Dataset Construction

In order to prepare the corpus, we have utilized scientific articles downloaded from digital libraries. This is so because it is observed that tables play a major role in depicting results and observations within scientific papers. To the process , we have downloaded 200 papers covering 20 different type of domains in computer science ,like Automatic Summary, Machine Learning , Machine Translation etc. The average number of sentences in each document is approximately 202, excluding title, author names and section headings. Table 1 shows the statistics of the corpus. The following steps illustrate the overview of the dataset construction steps.

3.1 Caption Sentence Extraction as Abstractive Summary

A well written caption can demonstrate the content of a table coherently. Therefore, we have written python scripts (python version 2.7) for extracting the captions for all the tables. A caption can be written in various formats depending on the domain and writing style. In order to deal with this variation, we have developed a method to differentiate caption sentences from other sentences in the document. We have observed from various papers, that a caption sentence consists of 4 parts. They are <*TABLE*> which refer to the word Table, followed by <*INTEGER*>, which is an integer that refers to the table number in the paper. The integer is followed by a <*DELIMETER*> which refers to the delimiter at the end of the sentence like "."or ": ". Finally we have <*TEXT*> which is the description of the table content. If a sentence follows this pattern, we label it as a caption sentence which then forms the abstractive summary content of that table.

3.2 Relevant Sentence Extraction as Extractive summary

Although a caption describes the content of a table quite elegantly, studies have shown that captions ,on their own, are insufficient in describing an element to a reader.To handle this issue, we have observed that any table is referenced at least once in the document. Thus to obtain a more comprehensive understanding of the table under consideration, we have extracted its reference text from the corresponding scientific document. Our first step was to segment the document text into sentences. For identifying relevant sentences, we have followed the same pattern as described for caption extraction, with the difference in the fact that the delimiter part is absent in such sentences. Moreover, when a table is referenced in the document, the sentences which are within a certain proximity of the reference sentence, are very useful in describing the context in which the table is being mentioned. Keeping this in mind , we have assigned scores to each sentence depending on its proximity level and distance to the reference sentence. If the distance is within a certain threshold length (+/-1), we have considered it as an important sentence and included it in the summary.

3.3 Data Annotation

Once the gold standard abstractive and extractive summaries are generated, an annotated dataset is constructed to create a structured, well defined and easy to use corpus. Two separate approaches were employed in order to construct a gold standard dataset capable of automatic evaluation and also to evaluate the efficiency of our system generated output as extractive and abstractive summaries. Firstly, an external annotator was employed, who is well versed in the field of computer science. This annotator was asked to manually identify the abstractive caption based summary and extractive summary only from the dataset. Secondly, our system generated an output feature file with additional features besides the summaries. These features are paper ID, table ID, no. of rows in a table, no. of rows in a table, row attributes, column attributes, and Table type (numeric, text or hybrid).

3.4 Evaluating the Quality of the System

The gold standard corpus consists of 200 papers and 499 tables.The following subsection discusses the evaluation process briefly.

3.4.1 Inter Annotator Agreement

To help us with our evaluation, a manual annotator A1 is employed. Annotator A1 and our system are then each arbitrarily provided with 100 separate documents each and instructed to identify both the extractive and abstractive summaries separately. For both, a score "1" was assigned to each of the sentences included in the summary and "0", for the one's which are not. In order to have an idea about the degree of agreement , we selected 100 papers, whose output are generated by A1 and another 100 papers whose output are generated by my proposed system. These papers are then interchanged and the annotators were asked to either agree (1) or disagree (0) with the other output.

Thus at the end we had 200 papers scored by both annotators. Out of this, 20 tables were selected, containing a total of 4040 summary sentences each, for extractive and abstractive summary. This is then used for measuring the agreement score between annotator A1 and the system generated output using **Cohen's Kappa coefficient** κ which is defined as

$$\kappa = \frac{Pr_a - Pr_e}{1 - Pr_e}, \qquad (1)$$

Where Pr_a is the observed proportion of full agreement between two annotators. In addition, Pr_e is the proportion expected by a chance and so indicates kind of random agreement between annotators.

The Cohen's Kappa agreement analysis provides κ = 0.83 and κ = 0.81 for extractive and abstractive summary agreement individually. Consequently, higher κ proves that the agreement is strong. The aim of this experiment was to evaluate how well the proposed method is able to identify the table content summary from the document.

3.4.2 Guidelines

The corpus contains separate folders for each of the 200 scientific papers of the computer science domain that we have processed.Table 1 gives a statistics of raw data. Within each folder, there are 6 separate files.

- CSV, which contains separate CSV files describing the content of each table in that paper.
- Annotation, which is the system generated feature file description for all tables in that paper.
- Document_PDF, which is the PDF version of the paper.
- Document_TXT, which contains the text version of the paper.
- Document_XML, which is the XML version of the paper.
- Summary, which contains the manually identified extractive and abstractive summary for each individual table of the paper.

Finally, a README file is included with the corpus, which describes each aspect of all the files.

4 System Design

We have proposed two models for generating templates that represent the extractive summary of a table. Our system produces a set of important terms, from each of the downloaded scientific papers. These terms are then used to generate an extractive summary of the tables contained in that paper. Finally, we have measured the similarity score between the generated templates and the gold standard summaries to evaluate the quality of summary in the dataset. The systems are described in the following subsection:

Paper Type	# Tables	Type: Text	Type: Numeric
Automatic Summary	50	21	29
Machine Learning	45	22	23
Machine Translation	55	19	36
NER	51	25	26
Question Answering	60	25	35
Sentiment Analysis	42	19	23
Speech Recognition	31	19	12
Text Classification	44	21	23
Text Segmentation	62	20	42
WSD	59	28	31
Total No.of Papers: 200			

Table 1: Statistics of the Corpus

4.1 TF-IDF Based System

4.1.1 Unigram Approach

These terms are selected by the following method:

–Initially a corpus is prepared from which gold standard extractive summaries for each table are extracted from that corpus.
–A set of unique words are collected from the gold standard dataset. Unique words are referred to as the words that are frequent or common in the reference summary.
–In the corpus there are 200 papers. For each table, in each paper, the TF-IDF score of all the terms are calculated excluding the stop words,non alpha-numeric characters and unnecessary punctuation. *TF* is the frequency of the term in that paper and *IDF* is the number of sentences in the paper where the term has occurred.
–Only those terms are considered that are within the set of unique words and belong to the highest scored terms for the template. These set of terms are referred to as Template for match(TS).
Each table can have multiple extractive summaries but there is only one TS for all the summaries of a particular table. So, we ranked them in order to see that with which extractive summary, the TS matched better. We have used Textual entailment method for ranking purpose.

4.1.2 Bigram Approach

A Bigram approach is also designed which takes into account a bigram instead of unigram. In this case, the TF-IDF score of a bigram is calculated in the document. But,here we have used only BLEU and ROUGE metrics for selecting terms in TS, whereas, we used Cosine Similarity, BLEU and ROUGE metrics in case of unigram. Also, all these scores have been used as a background knowledge in textual entailment which is used for ranking of the gold standard extractive summaries. In the evaluation section we have discussed which approach is the best among these two for generating extractive summaries.

4.2 Transition Point Based System

Transition Point(TP) is a frequency value that splits the vocabulary of a text into two sets of terms(low and high frequency). (Urbizagástegui, 1999) in thier paper, used the transition point(TP) to show its usefulness in text indexing. The mid-frequency terms are closely related to the conceptual content of a document.

4.2.1 Unigram Approach

A document$_i$ and its vocabulary V$_i$ =$\{(w_j,tf_i(w_j))|w_j\in D_i\}$,where $tf_i(w_j)=tf_{ij}$,let TP$_i$ be the transition point of D$_i$.A set of important terms which will represent the document D$_i$ may be calculated as follows:

$$R_i = \{w_j|((w_j, tf_ij) \in V_i), \\ (TP_i.(1-u) \leq tf_ij \leq TP_i.(1+u))\} \quad (2)$$

where u is a value in [0,1].Some experiments presented in (Urbizagástegui, 1999) have shown that u=0.4 is a good value for this threshold.TP is obtained using the following formula:

$$TP = \frac{-1+\sqrt{8 \times I_i + 1}}{2} \quad (3)$$

where I$_1$ represents the number of words with frequency equal to 1.We consider that terms whose frequencies are closer to TP,are important terms and hence will get a high weight for summarization.All other terms will get a weight close to zero.

4.2.2 Bigram Approach

The terms selected by the above method were enriched with the words which have similar characteristics. This was done using a co-occurrence bigrams based formula in (López et al., 2007). We have divided this bigram version of the system into three subsystems, viz. **Module-I**, **Module-II**, **Module-III**.

Formally, given a document $\mathrm{D_i}$ made up of only these terms selected by using the TP unigram approach($\mathrm{R_i}$), the new important terms for $\mathrm{D_i}$ will be obtained in different way for three subsystems. We have taken bigram of each document and calculated their TF score. TF score is calculated as number of times the bigram occurs in the document.

4.2.3 Module-I:Left Approach

Module-I considers TF score of the bigram, whose value is greater than one. In each of these bigram , if the terms that belongs to $\mathrm{R_i}$ is in right most position, then, we have considered the left term. This term is the new term that is included for $\mathrm{D_i}$. Formally, the new terms are generated according to the following expression:

$$\begin{aligned} R_i' = R_i \cup \{w' | (w_j \in R_i), \\ (v = w' \cdot w_j), (v \in D_i), (tf_i(v) > 1)\} \end{aligned} \tag{4}$$

4.2.4 Module-II:Right Approach

Module-II considers the right most term of the bigram when the terms in $\mathrm{R_i}$ is in left most position. Therefor, the new terms are obtained as follows:

$$\begin{aligned} R_i' = R_i \cup \{w' | (w_j \in R_i), \\ (v = w_j \cdot w'), (v \in D_i), (tf_i(v) > 1)\} \end{aligned} \tag{5}$$

4.2.5 Module-III:Left-Right Approach

Module-III is the integrated approach of Module-I and Module-II. Here, we have considered both left or right terms, whenever the terms in $\mathrm{D_i}$ is present in bigram. Formally,the new terms are obtained as follows:

$$\begin{aligned} R_i' = R_i \cup \{w' | (w_j \in R_i), \\ (v = w_j \cdot w' or v = w' \cdot w_j), (v \in D_i), \\ (tf_i(v) > 1)\} \end{aligned} \tag{6}$$

We only used a window of size one around each term of $\mathrm{R_i}$, and a minimum frequency of two for each bigram was required as condition to include new terms.

5 Evaluation

We have generated templates representing the extractive summary of a table using the above described systems. Then, we have ranked the templates in order to see which template has matched better with our gold standard summary. We have measured the similarity score between reference summary and templates using some standard metrics such as Cosine Similarity, ROUGE, BLEU.

6 Experiment and Results

In this work, we have proposed two systems, a TF-IDF based and a Transition point based, for generating extractive summaries.

In **TF-IDF based system** we have generated a Template for Matching(TS) with the highest TF-IDF scored terms. Now, the number of terms to be considered for TS solely depends on the result. Therefore, we experimented by taking variable number of such terms as shown in Table 2 and Table 3 .

Transition Point based system has two versions, the unigram and the bigram approach.The system based on unigrams, generates a set of terms whose frequency is close to the transition point. Similarity scores are then measured between the gold standard summary and system generated summary, using BLEU and ROUGE metrics. We have divided the bigram version of this system into three systems. We have measured their performance using the same similarity metrics and compared them with each other. We have also compared the unigram and bigram approach in the following section. The results are shown in Table 4.

7 Observation

In the experiment section we have mentioned that the experiments were done by changing the number of terms in Template for Matching (TS). It was observed that for smaller number of terms, the Cosine similarity and BLEU scores increased but there was a decrease in ROUGE score. If observed carefully, it can be seen that the top 10 terms give

# Terms	Cosine Similarity	BLEU	ROUGE		
			Precision	Recall	F-Measure
10	0.32	0.46	0.26	0.71	0.34
20	0.24	0.40	0.53	0.60	0.51
30	0.19	0.36	0.72	0.50	0.53

Table 2: Results obtained from TF-IDF unigram approach

# Terms	BLEU	ROUGE		
		Precision	Recall	F-Measure
10	0	0.003	0.009	0.004
20	0	0.003	0.009	0.004
30	0	0.002	0.1	0.003

Table 3: Results of TF-IDF bigram approach

Systems		BLEU	ROGUE		
			Precision	Recall	F-Measure
Unigram		0.044	0.36	0.47	0.08
Bigram	Module I	0.08	0.81	0.36	0.45
	Module II	0.11	0.79	0.39	0.46
	Module III	0.13	0.43	0.02	0.04

Table 4: Results obtained from Transition Point based system

Model	Approach	BLEU	ROGUE		
			Precision	Recall	F-Measure
TF-IDF	Unigram	0.46	0.26	0.71	0.34
	Bigram	0	0.003	0.009	0.004
TP	Unigram	0.044	0.36	0.047	0.08
	Bigram	0.13	0.43	0.02	0.46

Table 5: Comparison between TF-IDF and Transition Point

highest cosine similarity and BLEU scores. However, when the top 30 terms are considered, it was ROUGR which gave the highest score.

A comparison study has also been done between the proposed TF-IDF and Transition point based systems. The comparison is shown in Table 5 . We have considered only the best results obtained for each case. It is observed that in the TF-IDF unigram approach, BLEU score is better and in the Transition Point bigram approach, F-measure is better. But, It can be safely inferred that overall TF-IDF approach outperforms the Transition Point approach.

A set of unique words e.g. size,obtained,accuracy,lists,experiments etc. are collected from gold standard dataset for improving the quality of the generated template. We have tried to keep the words in template that belong to the set of unique words. In case of TF-IDF, we are able to include these unique words. Therefore, TF-IDF results are much better than Transition point based system.

8 Conclusion

In this work, we have presented our attempt to generate a gold standard corpus for table content summarization. While working it was found that the preparation of structured corpus was one of the greatest challenges. In the paper we have described how we have resolved all these challenges and prepared a corpus which is used for training,testing,evaluating a table summarization system. Moreover, we have also developed two models for the quality evaluation of our corpus. As a future scope, we plan to increase the size of the corpus as well as include semantic features along with lexical we are planning to design a semantic approach based system. .

References

Hsin-Hsi Chen, Shih-Chung Tsai, and Jin-He Tsai. 2000. Mining tables from large scale html texts. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 166–172. Association for Computational Linguistics.

René Arnulfo García-Hernández, Romyna Montiel, Yulia Ledeneva, Eréndira Rendón, Alexander Gelbukh, and Rafael Cruz. 2008. Text summarization by sentence extraction using unsupervised learning. In *Mexican International Conference on Artificial Intelligence*, pages 133–143. Springer.

Marek Hlavac. 2013. stargazer: Latex code and ascii text for well-formatted regression and summary statistics tables. *URL: http://CRAN. R-project. org/package= stargazer*.

Yulia Ledeneva, Alexander Gelbukh, and René Arnulfo García-Hernández. 2008a. Terms derived from frequent sequences for extractive text summarization. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 593–604. Springer.

Yulia Ledeneva, Alexander Gelbukh, and R García Hernandez. 2008b. Keeping maximal frequent sequences facilitates extractive summarization. *Research in Computing Science*, 34:163–174.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 17–24. Association for Computational Linguistics.

Ming-Ling Lo, Kun-Lung Wu, and Philip S Yu. 2000. Tabsum: A flexible and dynamic table summarization approach. In *Distributed Computing Systems, 2000. Proceedings. 20th International Conference on*, pages 628–635. IEEE.

Ming-Ling Lo, Kun-Lung Wu, and Philip Shi-lung Yu. 2003. Method and apparatus for dynamic and flexible table summarization. US Patent 6,523,040.

Franco Rojas López, Héctor Jiménez-Salazar, and David Pinto. 2007. A competitive term selection method for information retrieval. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 468–475. Springer.

Hwee Tou Ng, Chung Yong Lim, and Jessica Li Teng Koo. 1999. Learning to recognize tables in free text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 443–450. Association for Computational Linguistics.

Thanh Tam Nguyen, Quoc Viet Hung Nguyen, Matthias Weidlich, and Karl Aberer. 2015. Result selection and summarization for web table search. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 231–242. IEEE.

K Selçuk Candan, Huiping Cao, Yan Qi, and Maria Luisa Sapino. 2009. Alphasum: size-constrained table summarization using value lattices. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 96–107. ACM.

Ashwin Tengli, Yiming Yang, and Nian Li Ma. 2004. Learning table extraction from examples. In *Proceedings of the 20th international conference on Computational Linguistics*, page 987. Association for Computational Linguistics.

R Urbizagástegui. 1999. Las posibilidades de la ley de zipf en la indización automática. *Reporte de la Universidad de California Riverside*.

Yalin Wang and Jianying Hu. 2002. A machine learning based approach for table detection on the web. In *Proceedings of the 11th international conference on World Wide Web*, pages 242–250. ACM.