2016 IEEE Second International Conference on Big Data Computing Service and Applications

An Iterative Graph-based Generic Single and Multi Document Summarization Approach using Semantic Role Labeling and Wikipedia Concepts

Muhidin Mohamed, Mourad Oussalah
Department of Electronic, Electrical and Systems Engineering, School of Engineering
University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
Email: mam256@bham.ac.uk, m.oussalah@bham.ac.uk

Abstract— **This paper proposes an innovative graph-based text summarization model for generic single and multi-document summarization. The approach involves four unique processing stages: parsing sentences semantically using Semantic Role Labeling (SRL), grouping semantic arguments while matching semantic roles to Wikipedia concepts, constructing a weighted semantic graph for each document and linking its sentences (nodes) through the semantic relatedness of the Wikipedia concepts. An iterative ranking algorithm is then applied to the document graphs to extract the most important sentences deemed as the summary. The empirical evaluation of the proposed summarization model on a standard dataset from the Document Understanding Conference (DUC) showed the effectiveness of the approach which outperformed the baseline comparators in terms of ROUGE scores.**

Keywords- Semantic Role Labelling; Wikipedia concepts; iterative ranking algorithm; text summarization.

I. INTRODUCTION

Text Summarization (TS) aims to reduce the original text document into a short substitute summary which retains the most important facts of the source document. Today's increasing number of news sites, sent emails, customer product reviews, social media comments, tweets, blog posts and QA communities all contribute to the rapid growth of the already accelerating volume of textual information. This, in turn, renders the task of developing an effective text summarization system rather difficult. Therefore, advancing the research on TS is most needed than ever before due to the overwhelming growth of the Internet texts. Alongside this, the availability of full-fledged lexical knowledge sources and the powerful semantic analysis tools motivates an extensive exploration of knowledge-based summarization methods [1]. In this paper, we investigate a graph-based summarization approach using Semantic Role Labelling (SRL) for sentence level semantic parsing and the Wikipedia as a background knowledge source. The SRL is a semantic parsing method in Natural Language Processing (NLP) which identifies the semantic arguments associated with the predicate verbs of a sentence [2]. Figure 1 shows a semantically parsed example sentence using Lund Semantic Role Labeler[1]. In this case, the parser recognises three main semantic arguments; namely, the subject, the object and an indirect object, labelled as A0, A1 and A2, respectively. The use of the SRL and the Wikipedia concepts for text summarization is partially motivated by the findings of our previous work [3] where a Wikipedia-based measure for relating named entities was successfully applied to enhance the performance of a query-focussed text summarizer.

	John	finished	reading	the	book	and	gave	it	to	Mohamed	.
finish.02	A1		A1								
read.01	A0			A1							
give.01	A0							A2	A2		

Figure 1: An example sentence with SRL parsing.

The contributions of this paper are as follows. First, we avail SRL-based semantic representation of sentences to group similar arguments from each role-set and project them onto corresponding Wikipedia concepts. Second, we propose a weighted semantic document graph where each sentence is represented by the sub-nodes containing the concepts of its semantic arguments. The semantic relatedness between the Wikipedia concepts of the semantic arguments forms the edge-weights. Finally, the performance of our summarizer is empirically validated using the standard DUC2002 dataset. The rest of the paper is organized as follows. Section 2 covers a detailed explanation of the proposed summarization approach. Evaluation experiments of the summarizer are presented in Section 3 before drawing the paper conclusions in Section 4.

II. THE PROPOSED SUMMARIZATION MODEL

In a graph-based representation of textual documents, text units (e.g., words or sentences) form the nodes (vertices) of the graph whilst the associations between these units fill the position of the edges. In the context of summarization, the sentence similarities form the associations, if the nodes contain sentences. The use of graph-based algorithms for text summarisation has been widely explored in [4-7]. This paper extends the graph-based text summarization approach by exploiting Semantic Role Labelling and Wikipedia's rich concept structure to design an effective generic single and multi-document summarization model.

The proposed SRL Wikipedia graph based summarization model is illustrated in Figure 2. The process involves two stages. In the first stage, we perform two parallel processing tasks: data pre-processing and semantic parsing with SRL on the one hand, and constructing an inverted index file of Wikipedia concepts, on the other hand. The next stage deals with the core summarization tasks. In the pre-processing phase, we process the experimental dataset by converting the raw document texts into semantic units using basic NLP tasks, such as document segmentation, merging document sets (multi-documents), sentence tokenization, part-of-speech tagging, word stemming and the removal of stop words.

[1] http://barbar.cs.lth.se:8081/parse

978-1-5090-2251-9/16 $31.00 © 2016 IEEE
DOI 10.1109/BigDataService.2016.31

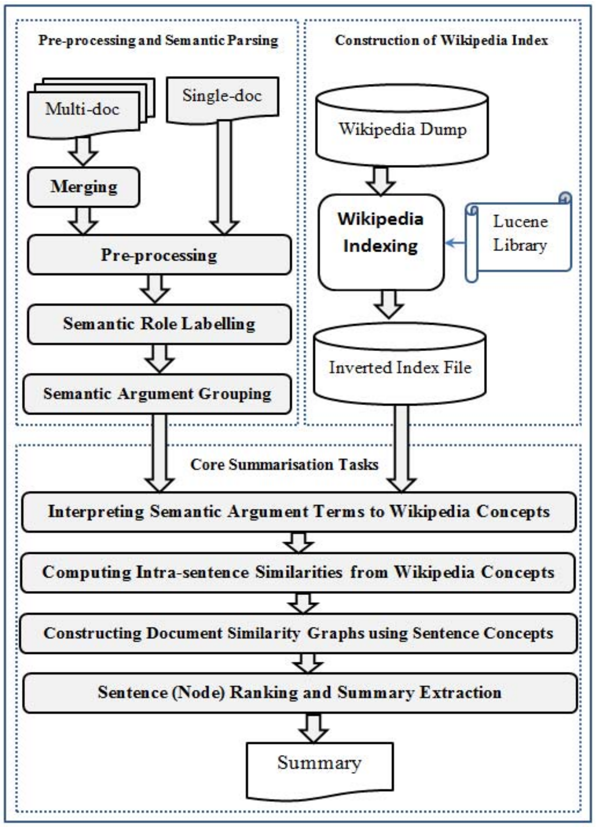

Figure 2: The SRL Wikipedia graph based summarization.

Next, we applied semantic parsing with SRL technique to identify the semantic frames and associated arguments. For instance, let S_1 and S_2 be two sentences consisting of the semantic frames f_1 and f_2, respectively. If $R_1 = \{r_1, r_2, \ldots, r_i\}$ and $R_2 = \{r_1, r_2, \ldots, r_j\}$ are the semantic role sets associated with f_1 and f_2, where i and j are the argument numbers in the semantic frames, we select the common roles, say, $R_c = \{r_1, r_2, \ldots, r_m\}$, co-occurring in both sentences, where m denotes the number of common roles. All other unshared semantic roles are discarded. This is because of the assertion that an accurate similarity can be captured by comparing the arguments corresponding to matching roles. The next step builds role-term vectors in which argument terms of similar semantic roles are grouped and linked to their modifiers.

This summarization model uses the Wikipedia database dump of 5th February 2015 to construct an inverted index file (see Figure 2). The inverted index file creates a mapping of argument terms to accommodating Wikipedia concepts using Explicit Semantic Analysis (ESA) method [8]. TF-IDF (term frequency- inverse document frequency) weighing, a very common weighting scheme in information retrieval (IR), is employed to quantify the term-to-concept associations.

A. Sentence Semantic Similarity

Sentence similarity scores are used to weight the links of the document similarity graphs. More specifically, we calculate the sentence similarity as the average semantic relatedness of the Wikipedia concept vectors representing the argument terms. Let $\{CV_{k1}, \ldots, CV_{ki}\}$ and $\{CV_{l1}, \ldots, CV_{li}\}$ be the concept vectors interpreted from the argument terms of the common roles between sentences k and l. The SRL Wikipedia-based similarity between sentence k and sentence l ($Sim_{Srl-Wp}(S_k, S_l)$) is calculated as the average role similarities (RSim) between the shared role sets as defined in expression (1) where i designates the shared semantic roles.

$$Sim_{Srl-Wp}(S_k, S_l) = \frac{1}{m}\sum_{i=1}^{m} RSim(CV_{ki}, CV_{li}) \qquad (1)$$

Here, $RSim(CV_{ki}, CV_{li})$ is computed using the concept vectors translated from the argument terms as in expression (2).

$$RSim(CV_{ki}, CV_{li}) = \frac{\sum_{j=1} wc_{jk} * wc_{jl}}{\sqrt{\sum_{j=1} wc_{jk}^2}\sqrt{\sum_{j=1} wc_{jl}^2}} \qquad (2)$$

where wc_{jk} $(resp.\ wc_{jl})$ represents the tf-idf weight of term j with respect to the corresponding concept of the argument role i of sentence k (resp. l).

B. Semantic Graph Representation of Documents

Each document is represented as a weighted undirected graph where the sentence concepts form the vertices and their semantic similarities weight the edges. More formally, let $G = (V, E, \alpha, \beta)$ be a weighed undirected graph with the set of vertices V representing sentence concept vectors and the set of edges $E \subseteq V$ linking the vertices. The parameters: $\alpha: V \rightarrow \Re_+$ and $\beta: E \rightarrow \Re_+$ are functions defining the vertex ranks and the edge weights respectively. In addition to the sentence-based graph representation, we used semantic links under sentence level. In other words, each sentence is modeled as a multi-node vertex using the Wikipedia concept vectors (CV) of the semantic argument terms, as shown in Figure 3. The semantic argument representation (A) and the sentence level (B) graphs are used for similarity computation and sentence ranking respectively. The edge-weights are formulated as per expression (3). For the single document summarization, the weights are measured in a slightly different way than for the multi-document summarization. For the former, the similarity between each sentence with the document title is considered in addition to the intra-sentence similarities. This is because, unlike multi-documents, each single document in the dataset has a unique title. Intuitively, having a high sentence similarity with the title indicates an additional importance of that given sentence.

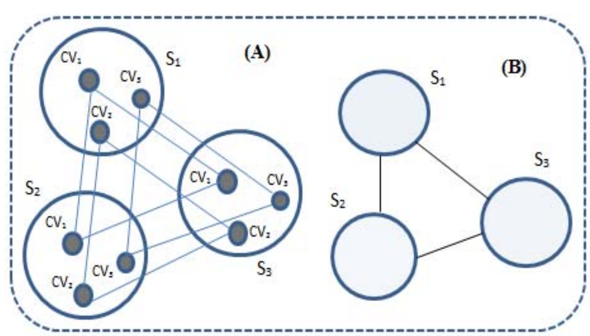

Figure 3: Semantic argument level representation (A) and sentence level (B) document similarity graphs.

$$Edge\ Weight\ (\beta) = \begin{cases} Sim_{Srl-Wp}(S_1, S_2) + TSim(S_1, S_2, T) & for\ SDS \\ Sim_{Srl-Wp}(S_1, S_2) & for\ MDS \end{cases} \quad (3)$$

Here, S_i denotes sentence i, T is the document title, SDS (resp. MDS) represents single document summarisation (resp. multi-document summarization) and $TSim(S_1, S_2, T)$ is the title similarity which is formulated in expression (4).

$$TSim(S_1, S_2, T) = 0.5 * \left(Sim_{Wp}(S_1, T) + Sim_{Wp}(S_2, T)\right) \quad (4)$$

The title-sentence similarity is calculated using Wikipedia concepts without semantic parsing due to the nature of most document titles which lack predicates and semantic frames. It is also worth noting that, in some rare cases, the sentences without predicate verbs are not included in the graph representation as will be seen in the example of Section C. This is because the SRL based semantic analysis cannot be applied to such sentences lacking semantic frames.

C. *Iterative Sentence Ranking for Summary Extraction*

We applied the PageRank algorithm [9] (expression 5) to the document similarity graphs for ranking and identifying the most important sentences [10].

$$\mathrm{PR}(\mathrm{p_i}) = \frac{1-\lambda}{\mathrm{N}} + \lambda * \sum_{\mathrm{p_j} \in \mathrm{In}(\mathrm{p_i})} \frac{\mathrm{PR}(\mathrm{p_j})}{\mathrm{Out}(\mathrm{p_j})} \quad (5)$$

where $\mathrm{PR}(\mathrm{p_i})$ is the Page Rank of the page $\mathrm{p_i}$, $\mathrm{In}(\mathrm{p_i})$ and $\mathrm{Out}(\mathrm{p_j})$ are the incoming and outgoing links for pages $\mathrm{p_i}$ and $\mathrm{p_j}$ respectively, N is the total number of pages, and λ is a damping factor which can be set between 0 and 1.

---

1 BFN Text Guangzhou, 19 Jun XINHUA -- Jiang Zemin, general secretary of the CPC Central Committee Political Bureau and chairman of the Central Military Commission, and Li Peng, premier of the State Council, are very much concerned about floods in Guangdong Province. 2 Recently, they repeatedly inquired about the flood situation in the Zhu Jiang valley, particularly that of Bei Jiang and Xi Jiang. 3 They expressed their deep concern for the people in the flood-hit areas, as well as extended their warm greetings to the vast number of cadres, officers, and men of the People's Liberation Army; armed police officers; and public security police who battle on the frontline against floods and provide disaster relief. 4 Jiang Zemin and Li Peng gave important directives for current flood prevention and disaster relief tasks in Guangdong. 5 They expressed the hope that under the leadership of the Guangdong provincial party committee and government, the Guangdong army and people would make concerted efforts in disaster relief; earnestly help flood victims solve their living problems; and go all out to battle floods to ensure the safety of the Bei Jiang dike, Guangzhou city, and the Zhu Jiang Delta.

Figure 4: A sample document to be summarized.

In the summarization context, we rank sentences instead of web pages; hence, sentences play the role of webpages. Similarly, intra-sentence semantic similarities substitute the incoming and outgoing links in the computation of sentence ranks. The rank of each sentence indicates its salience. This depends on the number and the strength of semantic links a sentence has with the rest of the sentences. Sentences with high similarities are more likely to be candidates for summary inclusion. To exemplify our reasoning, a short document of 5 sentences taken from the DUC2002 dataset is represented in Figure 4.

Parsing the document semantically reveals that the first sentence does not contain a predicate verb (or a semantic frame) and hence is excluded from further processing. This leaves the document with four sentences (2-5) to be summarized. Figure 5 shows the sentence similarity graph of the remaining four sentences indicating their edge weights and node ranks (the numbers in the square brackets). The document sentences are then ranked according to their importance as (5, 4, 3, 2) with the most and the least salient being the fifth and the second sentences respectively.

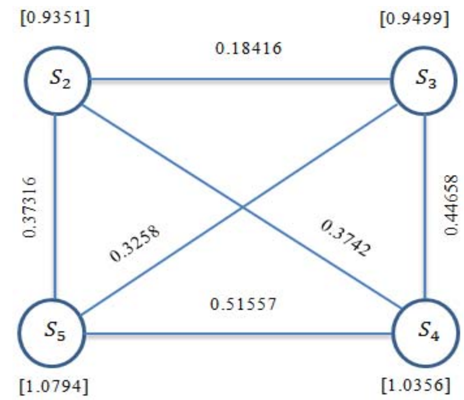

Figure 5: A sentence similarity graph for the document FBIS4-26327 with edge-weight and sentence rank scores after 20 iterations.

III. EXPERIMENTS

A. *Evaluation*

For the purpose of testing and validation, we used 21 clusters of 160 documents from the DUC2002 dataset. The entire DUC2002 collection contains 60 sets of about 10 documents. Each document has an abstract model summary of 100 words. We also employed the Recall Oriented Understudy for Gisting Evaluation (ROUGE) [11], which is the most widely used official evaluation tool in text summarization. As in expression (6), the ROUGE determines the quality of a system summary by comparing its text to an ideal human summary (also known as a reference summary).

$$ROUGE - N = \frac{\sum_{S \in RS} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in RS} \sum_{gram_n \in S} Count(gram_n)} \quad (6)$$

Here, N is the length of the $gram_n$, $Count(gram_n)$ is the number of n-grams ($gram_n$) in the reference summary while $Count_{match}(gram_n)$ is the maximum number of n-gram matches in the system summary and the collection of the reference summaries (RS). A $gram_n$ refers to a sequence of n words, e.g., a two-word sequence is called a bigram.

B. *Results and Discussion*

The PageRank algorithm is iteratively run on the document similarity graphs (see Section 2-C) until it converges. A sentence with a high semantic similarity score and linked with many other sentences are ranked higher. Next, the highest ranked sentences of each document are extracted as a summary. In most cases, our experimental results proved that the algorithm converges before reaching the 20th iteration. Following the DUC guidelines, we extracted 100-word and 200-word summaries for SDS and MDS respectively. Tables 1-2 show the quality of the system summaries produced for the SDS and MDS in terms of the average ROUGE recall scores at a 95% confidence interval.

Table 1: Results of the SRL Wikipedia graph based SDS.

	Recall	Precision	F-measure
ROUGE-1	0.5037	0.4305	0.4623
ROUGE-2	0.2353	0.2005	0.2156
ROUGE-L	0.3345	0.2857	0.3069
ROUGE-SU4	0.2537	0.2156	0.2321

Table 2: Results of the SRL Wikipedia graph based MDS.

	ROUGE-1	ROUGE-2	ROUGE-SU4
Recall	0.4743	0.2123	0.2455
Precision	0.4267	0.1902	0.2199
F-Measure	0.4489	0.2005	0.2318

Besides summarizing with the proposed SRL Wikipedia graph based system, we extracted document summaries of the same dataset with Microsoft Word Summarizer, which we used as a benchmark method. The Microsoft Word Summarizer is widely used in the related studies [12-13] as a benchmark method for summarization. In addition, the best performing system of the related DUC competition, labelled as System 19, is employed as another comparator. The bar-charts (A) and (B) in Figure 6 show the comparison of our results and those from the two comparators for the SDS and MDS tasks. The competency of the proposed SRL Wikipedia

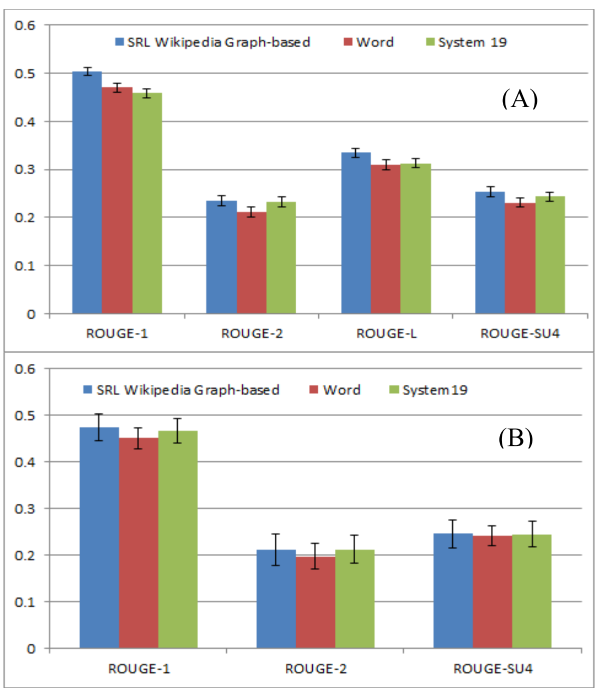

Figure 6: Comparing the ROUGE scores of the proposed SRL Wikipedia graph-based summarization model, the MS Word Summarizer, and the top related DUC system for single (A) and multi-document (B) summarization tasks.

graph based summarization is shown in Figure 6 where it outperforms both benchmark methods with variations in all ROUGE measures. The results indicate the advantages of the SRL technique and the Wikipedia concept structure for text summarization.

IV. Conclusion

This paper proposes a graph-based single document and multi-document summarization approach. The summarizer benefits from the SRL method and the Wikipedia concepts. The empirical evaluation of our summarization method on the standard DUC2002 dataset discloses that it improves the summary quality as compared to benchmark methods. This shows the effectiveness of the semantic arguments mapped to the Wikipedia concepts and the use of concept relations under the sentence level for computing sentence similarities.

References

[1] Nenkova, Ani, Sameer Maskey, and Yang Liu. "Automatic summarization."Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics, 2011.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR), 2011.

[3] Mohamed, Muhidin A., and Mourad Oussalah. "Similarity-Based Query-Focused Multi-document Summarization Using Crowdsourced and Manually-built Lexical-Semantic Resources." *Trustcom/BigDataSE/ISPA, 2015 IEEE*. Vol. 2. IEEE, 2015.

[4] Erkan, Günes, and Dragomir R. Radev. "LexRank: graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.

[5] Wei, Furu, et al. "A document-sensitive graph model for multi-document summarization." Knowledge and information systems 22.2 (2010): 245-259.

[6] Canhasi, Ercan, and Igor Kononenko. "Semantic role frames graph-based multidocument sumarization." SiKDD'11 (2011).

[7] Sankarasubramaniam, Yogesh, Krishnan Ramanathan, and Subhankar Ghosh. "Text summarization using Wikipedia." Information Processing & Management 50.3 (2014): 443-461.

[8] Gabrilovich, Evgeniy, and Shaul Markovitch. "Wikipedia-based semantic interpretation for natural language processing." Journal of Artificial Intelligence Research (2009): 443-498.

[9] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." Computer networks 56.18 (2012): 3825-3833.

[10] Mihalcea, Rada, and Paul Tarau. "Multi-document Summarization with iterative graph-based algorithms." (2005).

[11] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries."Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004.998.

[12] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm diversity hybrid model for text summarization," Information processing & management, vol. 46, pp. 571-588, 2010.

[13] S. Xu, H. Jiang, and F. Lau, "User-oriented document summarization through vision-based eye-tracking," in *Proceedings of the 14th international conference on Intelligent user interfaces*, 2009, pp. 7-16.