Information Processing and Management 56 (2019) 1356–1372

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis[☆]

Muhidin Mohamed[*,a], Mourad Oussalah[b]

[a] Computer Science, School of Engineering and Applied Sciences, Aston Universty, Aston Triangle, Birmingham B4 7ET, UK
[b] Centre for Ubiquitous Computing, Faculty of Information Technology Computer Science, University of Oulu, P.O. Box 4500, 90014 Finland

ARTICLE INFO

Keywords:
Text summarization
Semantic role labeling
Wikipedia concepts
Concept graphs
Semantic similarity
Iterative ranking algorithm

ABSTRACT

Automatic text summarization attempts to provide an effective solution to today's unprecedented growth of textual data. This paper proposes an innovative graph-based text summarization framework for generic single and multi document summarization. The summarizer benefits from two well-established text semantic representation techniques; Semantic Role Labelling (SRL) and Explicit Semantic Analysis (ESA) as well as the constantly evolving collective human knowledge in Wikipedia. The SRL is used to achieve sentence semantic parsing whose word tokens are represented as a vector of weighted Wikipedia concepts using ESA method. The essence of the developed framework is to construct a unique concept graph representation underpinned by semantic role-based multi-node (under sentence level) vertices for summarization. We have empirically evaluated the summarization system using the standard publicly available dataset from Document Understanding Conference 2002 (DUC 2002). Experimental results indicate that the proposed summarizer outperforms all state-of-the-art related comparators in the single document summarization based on the ROUGE-1 and ROUGE-2 measures, while also ranking second in the ROUGE-1 and ROUGE-SU4 scores for the multi-document summarization. On the other hand, the testing also demonstrates the scalability of the system, i.e., varying the evaluation data size is shown to have little impact on the summarizer performance, particularly for the single document summarization task. In a nutshell, the findings demonstrate the power of the role-based and vectorial semantic representation when combined with the crowd-sourced knowledge base in Wikipedia.

1. Introduction

Text Summarization (TS) is a Natural Language Processing (NLP) task aimed to reduce original text documents to a short substitute summary which retains the most important facts of the source document. Summarization can be achieved by either extracting a group of sentences from original source document (s) and concatenating them, called *extractive summarization*, or generating a novel summary text representing its main gist, referred to as *abstractive summarization*. From an *input* perspective, a summary can either be sourced from one document through a process called *single-document summarization*, or from a collection of related documents which is known as *multi-document summarization*. Depending on the desired content, a summary is either a *query-focused* (tailored to a user query), or a *topic-focused* (containing document gist).

[☆] Most of this work was done while a PhD student at the Uiversity of Birmingahm.

[*] Corresponding author.
E-mail addresses: mam256@alumni.bham.ac.uk, m.mohamed10@aston.ac.uk (M. Mohamed).

https://doi.org/10.1016/j.ipm.2019.04.003
Received 10 August 2018; Received in revised form 20 February 2019; Accepted 12 April 2019
Available online 15 April 2019

Today's increasing number of news sites, sent emails, customer product/service/travel reviews, user-generated social media content and QA communities all contribute to the rapid growth of already accelerating volume of textual information. The amount of information indexed in the Internet is estimated to be over 4.5 billion pages.[1] However, as the volume of generated structured and unstructured text increases exponentially, it renders the task of developing effective text summarization systems rather appealing. Therefore, advancing the research on TS is most needed than ever before due to the overwhelming growth of the Internet texts (Gambhir & Gupta, 2017). Alongside this, the availability of full-fledged lexical knowledge sources and the powerful text semantic analysis tools inspire an extensive exploration of semantic & knowledge based summarization methods. This is also motivated by the assertion that using semantic knowledge holds the potential for further improvements in text summarization research and therefore needs more research investigation (Nenkova, Maskey, & Liu, 2011).

In this paper, we investigate a graph-based summarization approach using semantic role labelling for sentence level semantic parsing and Wikipedia as an external knowledge source. The SRL is a semantic parsing technique in NLP which identifies the semantic arguments associated with the predicate verbs of a sentence. It classifies the semantic roles of syntactic arguments within a given frame of the sentence and with respect to the predicate. The use of the proposed SRL-ESA Wikipedia graph-based generic summarization model is motivated by the following: (1) the successful application of Wikipedia-based metric to the tasks of named entity semantic relatedness and query-focused summarization in our previous work (Mohamed, 2016; Mohamed & Oussalah, 2015); (2) the high lexical coverage of Wikipedia which led to its popularity as a reliable lexical resource for different NLP tasks such as semantic representation (Saif, Omar, Ab Aziz, Zainodin, & Salim, 2017), word semantic similarity (Jiang, Zhang, Tang, & Nie, 2015) and text classification (Wang, Hu, Zeng, & Chen, 2009); (3) the status quo in which current knowledge-based summarization methods underestimate the importance of sentence syntactic order and semantic roles, consequently leading to poor scoring functions for summary extraction; (4) the computation of word similarities in isolation from the surrounding context, thus ignoring significant semantic information conveyed by these words if associated with their roles. The chief contributions of this paper are:

1. First, we used semantic role labelling to build a semantic representation of documents and then paired matching semantic roles for any two compared sentences before mapping them onto their corresponding concepts in Wikipedia.
2. Next, we proposed a weighted semantic graph where each sentence is modeled as a *multi-node vertex* containing the Wikipedia concepts of its semantic arguments.
3. Finally, we implemented a single & multi-document summariser using the above mentioned semantic graph representation and empirically evaluated its performance using the standard DUC 2002 dataset.

The rest of the paper is organized as follows. Sections 2 and 3 cover the research objective and related works, respectively. Section 4 gives a brief introduction to the applied semantic analysis methods whereas Section 5 details the proposed summarization approach. Next, experiments for evaluating the system are presented in Section 6 before drawing paper conclusions in Section 7.

2. Research objective

In this study, we combine four important components to model an extractive text summarizer: text semantic representation, effective concept-based sentence similarity measure, semantic role-based multi-node concept graph representation, and sentence ranking algorithm. The goal is to investigate the aggregate effect of the aforementioned four factors on the performance of topic-focused single and multi-document summarization. Our rationale behind this is to integrate the advantages of well-established text semantic analysis and representation techniques and the vast human knowledge encoded in the Wikipedia database.

The use of semantic parsing and word semantic roles is to address the problem of greedy word pairing approaches when computing sentence similarity. To achieve that, semantic argument terms sitting the same role are grouped and mapped onto a vector of Wikipedia concepts. Wikipedia-mined concept vectors representing semantic arguments are then used to form sub-nodes of each sentence vertex in the document concept graph. We provide a detailed technical description of the proposed summarization system throughout the article and illustrate its functionality through a working example. We also present an experimental evaluation to highlight the effectiveness of the method.

3. Related works

Scoring document contents (e.g., words, phrases, sentences) is the most common method used in automatic extractive text summarization. The majority of today's implemented extractive summarizers adopt sentence scoring or graph-based sentence ranking. Prevailing sentence scoring and selection techniques in text summarization include graph-based representation (Oliveira et al., 2016), text semantic analysis (Mohamed, 2016), semantic similarity (Jiang et al., 2015), sentence clustering (Alguliyev, Aliguliyev, Isazade, Abdi, & Idris, 2018), fuzzy reasoning (Binwahlan, Salim, & Suanmali, 2010; Kumar, Salim, Abuobieda, & Albaham, 2014), sentence regression (Ren, Wei, Zhumin, Jun, & Zhou, 2016) and differential evolution (Abuobieda, Salim, Kumar, & Osman, 2013; Alguliyev, Aliguliyev, & Isazade, 2013). Several related works conducted a comparative study on a range of sentence scoring methods by examining the performance of their combinations for text summarization (Ferreira et al., 2013; Oliveira et al., 2016).

[1] http://www.worldwidewebsize.com/.

Two powerfull but often underused semantic analysis tools for text summarization are the SRL and ESA methods. A few related works have independently utilised SRL for text summarisation. Khan, Salim, and Kumar (2015) advocated a feature-based approach in associated with SRL where predicate argument structures were used to represent source documents. And Suanmali, Salim, and Binwahlan (2010) combined statistical and SRL-based features, whereas authors in Jha and Anas (2018) and Aksoy, Bugdayci, Gur, Uysal, and Can (2009) used semantic frame and argument frequencies to identify key sentences. Some other studies used SRL together with an iterative graph-based ranking algorithm for text summarization. For instance, the work of Canhasi and Kononenko (2011) who introduced a multilayered document similarity graph linking sentence semantic frames. Likewise, Yan and Wan (2014) used SRL tuples as additional nodes in a multi-level graph representation and treated them as independent units for sentence ranking. Moreover, the application of ESA to text summarisation is still in its infancy. On this subject, Sankarasubramaniam, Ramanathan, and Ghosh (2014) suggested a Wikipedia-based text summarisation algorithm using a bipartite sentence concept graphs to rank document sentences according to their concepts. In a more feature-based fashion, Zhou, Guo, Ren, and Yu (2010) applied ESA to query-focussed text summarisation and integrated an ESA-based technique and traditional sentence features to score document sentences using machine learning algorithms.

In addition, graph-based representations are some of the most prevalent text analysis methods and have shown their effectiveness for text summarization (Abdi, Shamsuddin, & Aliguliyev, 2018; Azadani, Ghadiri, & Davoodijam, 2018; Canhasi & Kononenko, 2011; 2014; Erkan & Radev, 2004; Mihalcea & Tarau, 2004; Wan, 2010; Wei, Li, Lu, & He, 2010). The conventional way of graph-based summarization uses document sentences as vertices, known as sentence-based document graphs. Erkan and Radev (2004) proposed one of the most popular sentence-based graph representations for summarisation. Their LexRank algorithm is based on the eigenvector centrality concept. Similarly, Mihalcea and Tarau (2004) presented TextRank, another graph-based ranking method constructed using content overlap. Both LexRank and TextRank are derivatives of the seminal PageRank algorithm (Brin & Page, 2012). Some graph-based approaches have cross-linked different levels of text granularities particularly tailored for multi-document summarization (Canhasi & Kononenko, 2011; Wan, 2010; Wei et al., 2010). In this way, Canhasi and Kononenko (2014) used three-layer graph representation consisting of terms, sentences, and document vertices, and linked them via term-sentence and sentence-document links on top of the conventional sentence similarity graphs. Wei et al. (2010) and Wan (2010) considered the influence of global information from the document clusters on local sentence evaluation. Contrary to the traditional way of representing source text units, concept graphs have also been emerging as alternative graph representation of the source texts (Azadani et al., 2018; Lloret & Palomar, 2012; Morales, Esteban, & Gervás, 2008; Sankarasubramaniam et al., 2014). Of this, Azadani et al. (2018) and Morales et al. (2008) have both adapted such a method by modelling a biomedical summarization algorithm on concept graphs. Also, Zhuge (2016) proposed a multi-dimensional summarization methodology to summarize various objects (including texts, pictures and videos) from multiple dimensions. The effectiveness of this methodology has been evaluated for text summarization (Sun & Zhuge, 2018). Concept graph modelling proved some success particularly in domain-specific areas such as biomedical and news summarizations (Lloret & Palomar, 2012).

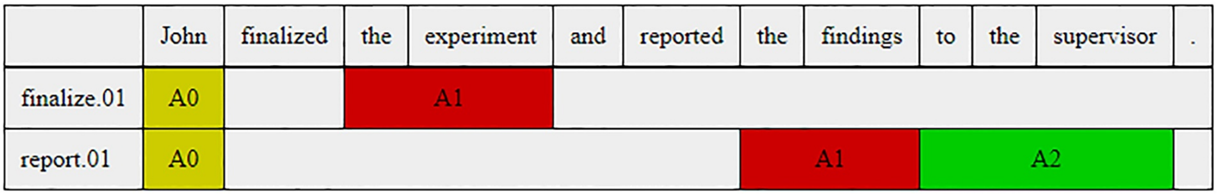
While the advantages of semantic parsing for text summarisation have been highlighted in previous semantic-based works (Jha & Anas, 2018; Khan et al., 2015; Yan & Wan, 2014), the uniqueness of our approach is that it combines the strengths of leveraging text semantic analysis and representation techniques with a high coverage crowd-sourced human knowledge. Also, the distinction between the current SRL-ESA Wikipedia based summarisation model and the preceding graph-based methods is the use of under-sentence semantic argument links in the construction of document concept graphs, which provides an edge to its competitors. This is because, intuitively, pairing matching semantic roles captures more semantics than applying indiscriminate word pairing greedily. Thus, realizing the strengths of world knowledge and semantic analysis, our approach adapts both SRL and ESA techniques for extractive text summarisation underpinned with the encyclopedic knowledge in Wikipedia.

4. Used semantic analysis techniques

4.1. Semantic role labeling (SRL)

SRL is a technique for sentence level semantic analysis. It segments the text and identifies the semantic role of each syntactic constituent word with reference to the main predicate verb in a sentence. On the other hand, semantic roles are the basic units of a semantic frame which is a collection of facts that specify *“characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it”* (Allan, 2001). Relations between semantic frames and word meanings, as encoded in FrameNet lexical database (Baker, Fillmore, & Lowe, 1998), represent the core of Frame Semantics theory (Fillmore & Baker, 2010). PropBank, another relevant resource, houses a large corpus of human annotated predicate-argument relations added to the syntactic trees of the Penn Treebank (Gildea & Jurafsky, 2002). The basic concept of Frame Semantics is that word meanings must be described in relation to semantic frames.

Linked with the above, sentence semantic parsing is a fundamental task that has a large number of immediate NLP applications including plagiarism detection (Osman, 2012) and text summarization (Suanmali et al., 2010). With the help of human annotated resources such as ProbBank (Palmer, Gildea, & Kingsbury, 2005) and FrameNet (Baker et al., 1998), the development of automatic systems for the identification of semantic roles is a well investigated current research topic in NLP. One of the seminal works about building automatic semantic role labellers was proposed by Gildea and Jurafsky (2002). Their system is based on statistical classifier trained on hand-annotated dataset from FrameNet. Recently, Collobert et al. (2011) proposed a unified neural network architecture and learning algorithm which was applied to different NLP tasks such as part-of-speech tagging, chunking, named entity recognition, and semantic role labelling. Their algorithm learns internal data representations using vast amounts of mostly un-annotated training



	John	finalized	the	experiment	and	reported	the	findings	to	the	supervisor	.
finalize.01	A0		A1									
report.01	A0						A1		A2			

Fig. 1. Example 1 semantically parsed with SRL.

data. They built an open-source software called SENNA which we used for the prediction of semantic roles in our work. One of the attractive features of this tagging system is its good performance in terms of its speed and the minimal computational requirements.

In a nutshell, the primary goal of SRL is to single out all component words that fill semantic roles for a predicate verb and then assign them the corresponding semantic role tags. It is usually stated that SRL answers the question of basic event structures such as *who did what to whom when where* and *why*. Example 1 illustrates how SRL works.

Example 1. John finalized the experiment and reported the findings to the supervisor.

Fig. 1 shows the sentence in Example 1 semantically parsed with the Lund Semantic Role Labeler.[2] The semantic parser recognises the predicate verbs and their associated arguments. Core SRL arguments include Agent (aka subject), Theme (aka direct object), Instrument, among others. They also include adjunctive arguments indicating Locative, Negation, Temporal, Purpose, Manner, Extent, Cause, etc. The list of semantic role arguments and associated tags are given in Table 1. Fig. 1 indicates that the example sentence has two verbs: *finalized* and *reported*. The labels A0, A1 and A2 in the figure indicate the subject, object and indirect object of the respective verb, whilst role-sets of the predicate verbs *finalized*, and *reported* are listed in Table 2. The hyphen (-) in the table indicates that the predicate lacks this argument. One can note that the subject *John* is a common agent for both verbs.

4.2. Explicit semantic analysis (ESA)

ESA is a Wikipedia-based technique for computing text semantic relatedness proposed by Gabrilovich and Markovitch (2009), and has been used for various other NLP tasks such as text categorisation (Gabrilovich & Markovitch, 2006) and information retrieval (Egozi, Markovitch, & Gabrilovich, 2011). The ESA procedure maps text snippets to a vector space containing Wikipedia-derived concepts. The technique assumes that Wikipedia articles represent natural language concepts and, hence, mapping text fragments to their accommodating concepts is perceived as a representation of the text meaning. Formally speaking, ESA constructs an inverted index from the Wikipedia database and uses that to represent input texts by building ordered and weighted Wikipedia concepts. This is done by iterating over each token of a text to be interpreted. The actual computation of the text semantic relatedness is then performed by comparing translated vectors of two texts using cosine similarity. Fig. 2 demonstrates the explicit semantic analysis process. For two natural language fragments to be compared the semantic interpreter iterates over each word of every text, retrieves its corresponding entry from the inverted index, and represents the word by the retrieved vector of concepts weighted by their TF-IDF scores.

More formally, if $T = \{w_i\}$ is the input text, $\overrightarrow{K}_{w_i}$ is the inverted index entry for word w_i, where K_{w_i} represents the strength of association for w_i with the Wikipedia concept set $C = \{c_1, c_2, ..., c_N\}$, then the semantic interpretation for T is the vector $V = \{v_1, v_2, ..., v_N\}$. Each element in V quantifies the association of the corresponding concept c_j to the text T, and is defined as $\sum_{w_i \in T} tf.\, idf_{w_i} {*} k_{w_i}$. The TF-IDF (term frequency- inverse document frequency) is one of commonest weighting schemes in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999). It calculates the weight of a word as per expression (1).

$$tf.\, idf\,(w,\, d) = tf_{w,d}\, .\, log\frac{N}{n_w} \tag{1}$$

where $tf_{w,\, d}$ is the frequency of word w in document d (Wikipedia article), n_w is the number of documents in which w occurs, and N is the number of documents in the text collection (size of English Wikipedia articles in our work). Once the text T is mapped onto its corresponding Wikipedia concepts vector, the final stage of the ESA process is to compute the semantic relatedness. In other words, if T_1 and T_2 are two text fragments, their semantic relatedness, $SemRel(T_1,\, T_2)$, is computed by comparing their respective vectors; V_1 and V_2 as in expression (2).

$$SemRel(T_1,\, T_2) = \frac{V_1.\, V_2}{\|V_1\|\|V_2\|} \tag{2}$$

5. Proposed SRL-ESA based text summarization model

To our knowledge, this work is the first study that combines Semantic Role Labeling and Wikipedia-based explicit semantic

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

Table 1
Semantic role arguments and associated labels.

Core arguments		Non-core arguments	
Label	Modifier	Label	Modifier
V	Verb	AM-DIR	Direction
A0	Subject	AM-ADV	Adverb
A1	Object	AM-LOC	Location
A2	Indirect Object	AM-TMP	Temporal marker
A3	Start Point	AM-MNR	Manner
A4	End Point	AM-DIS	Discourse marker
A5	Direction	AM-PRP	Purpose
–	–	AM-NEG	Negation
–	–	AM-EXT	Extent
–	–	AM-PNC	Proper noun

Table 2
Verb-argument pairs for the example in Fig. 1.

Verbs\Arguments	A0	A1	A2
Finalize	John	The experiment	–
Report	John	The findings	To the supervisor

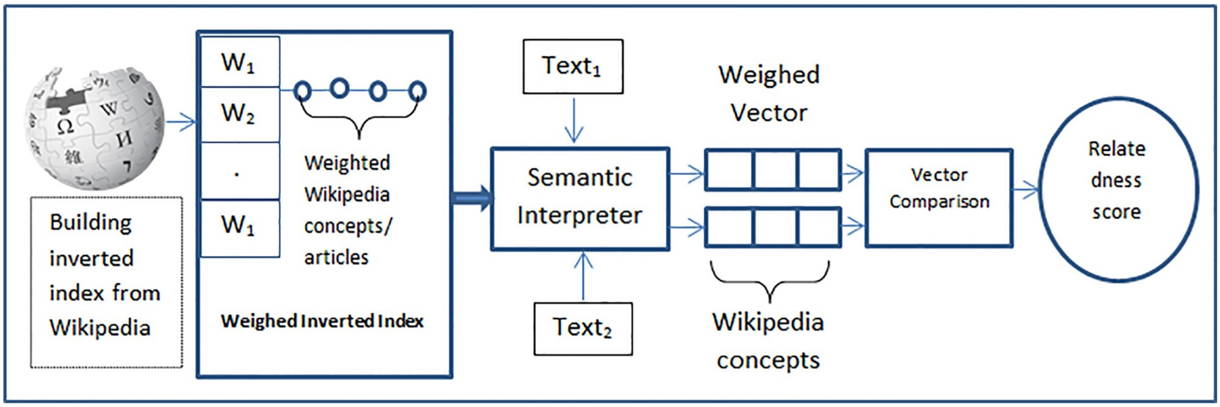


Fig. 2. Explicit semantic analysis.

analysis for text summarisation. It improves the construction of document similarity graphs for graph-based text summarisation. Fig. 3 shows a block diagram of the proposed text summarization model. The process involves two principle stages. In the first stage, we perform two parallel processing tasks: the experimental data pre-processing followed by a semantic parsing with SRL and argument grouping on the one hand, and constructing an inverted index database of Wikipedia concepts on the other hand. In addition, summarizing multi-documents involves additional step in which documents of each cluster are merged to form a single cluster document, as described in Section 5.1. In the pre-processing step, we processed the experimental dataset by converting the raw document texts to semantic linguistic units using some basic NLP tasks including document segmentation, sentence tokenization, part-of-speech tagging, word stemming and the removal of stop words. Document texts are parsed with semantic role labeling to identify the semantic frames and associated arguments. This is followed by semantic argument grouping in which all argument terms of similar semantic roles are collected and linked to their modifiers. The proposal makes use of Wikipedia database dump to construct an inverted index file (top right of Fig. 3). The inverted index file is created with the aid of Appache Lucene Library[3] to provide a mapping of argument terms to corresponding Wikipedia concepts using the ESA approach (Gabrilovich & Markovitch, 2009) as explained in Section 4.2. The next stage deals with core summarization tasks, as will be detailed in the following sections.

5.1. Merging multi-document clusters

In multi-document summarization, a single summary is sought from across many documents that describe the same topic. These

[3] http://lucene.apache.org.

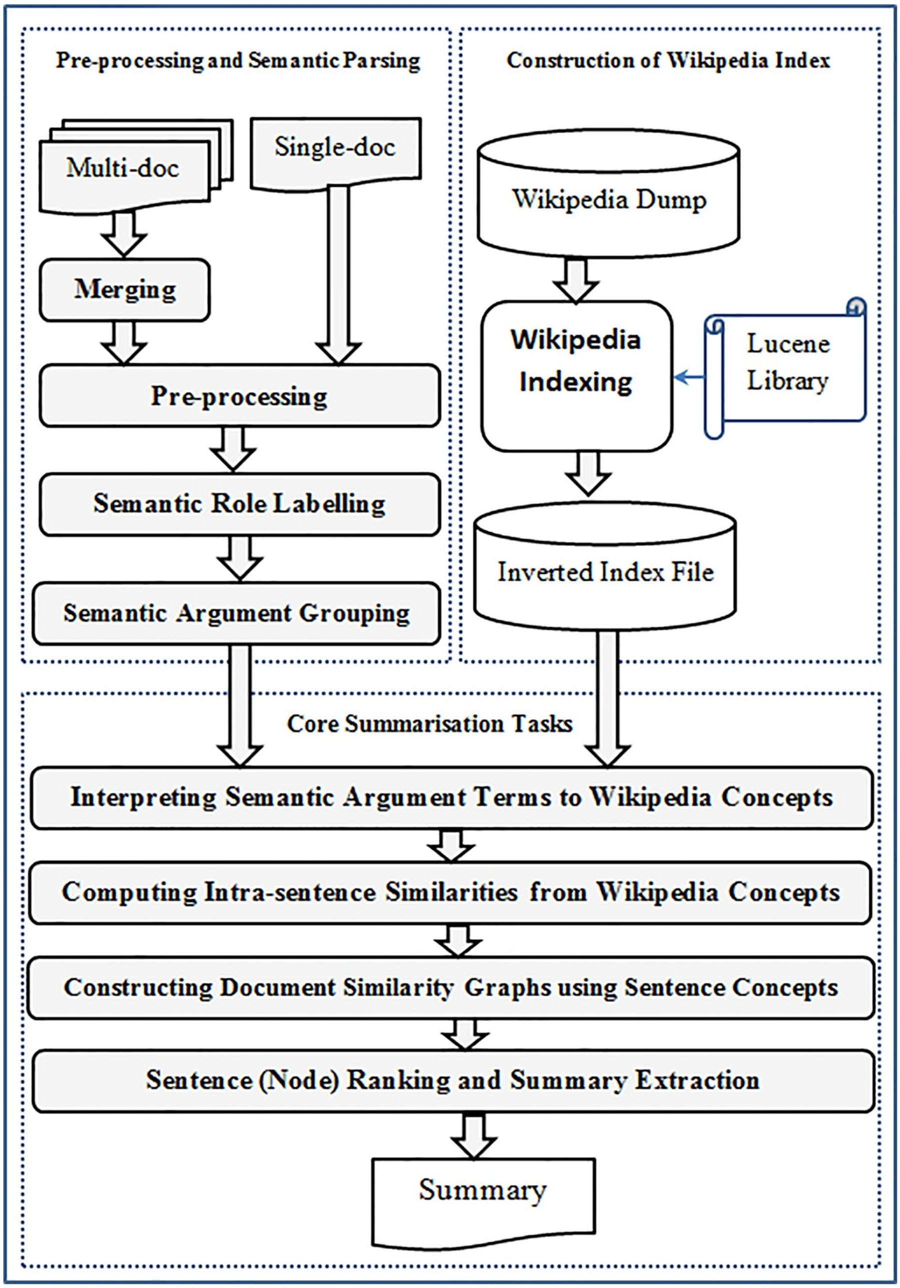

Pre-processing and Semantic Parsing
Multi-doc
Single-doc
Merging
Pre-processing
Semantic Role Labelling
Semantic Argument Grouping
Construction of Wikipedia Index
Wikipedia Dump
Wikipedia Indexing
Lucene Library
Inverted Index File
Core Summarisation Tasks
Interpreting Semantic Argument Terms to Wikipedia Concepts
Computing Intra-sentence Similarities from Wikipedia Concepts
Constructing Document Similarity Graphs using Sentence Concepts
Sentence (Node) Ranking and Summary Extraction
Summary

Fig. 3. SRL-ESA graph-based text summarization system.

documents, which are written by different authors, are normally taken from different news sources. Unlike single document summarisation, the process of summarising a collection of related documents poses a number of other challenges including a high degree of redundancy, which conceivably results from merging multiple descriptions of the same topic, and the ordering of the extracted summary sentences. To reduce redundancy, different summarization approaches used different methods such as measuring sentence similarity (Alguliyev et al., 2013; Mosa, Anwar, & Hamouda, 2019), using the seminal MMR algorithm and its derivatives (Binwahlan et al., 2010; Carbonell & Jade, 1998; Gambhir & Gupta, 2017), and exploiting clustering algorithms (Ferreira et al., 2014). Besides, sentence ordering remains a less studied problem in MDS. On this subject, Bollegala, Okazaki, and Ishizuka (2010) combined four criteria (*chronology, topical-closeness, precedence, and succession*) to develop a sentence ordering approach for multi document summarization. In this work, we have designed a pre-processing stage to mitigate these challenges. Firstly, related documents of each cluster to be summarised are merged together to form a single cluster document while arranging the entire text in the order of the source documents' timeline. We then iteratively removed similar sentences to exclude repeated content. This is done by finding the similarity of each sentence with the rest of the cluster sentences and removing those with high similarity scores. This produces a

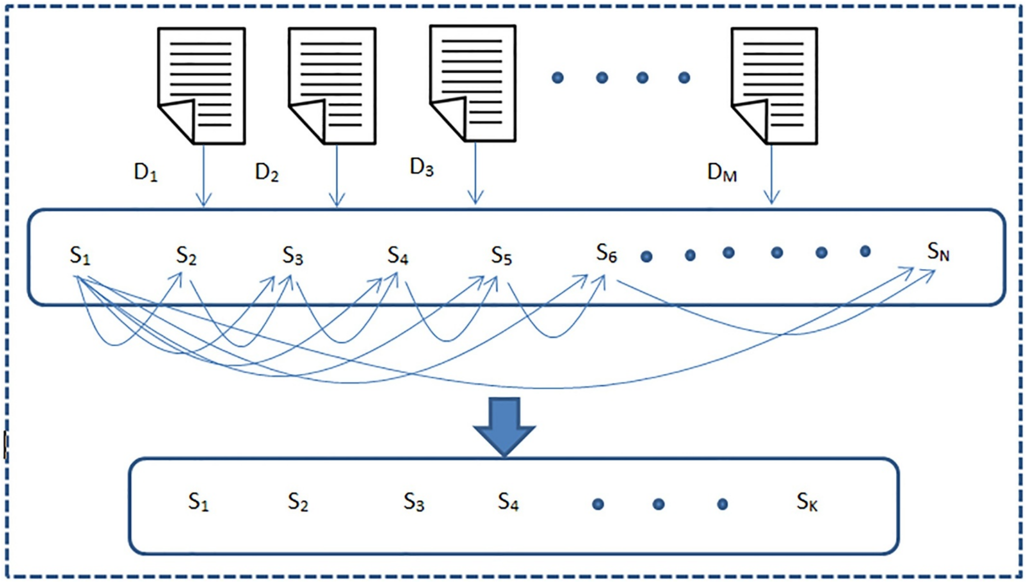


Fig. 4. Merging cluster documents with redundancy removal.

unified cluster document with minimized information repetition.

More formally, if $C = \{D_1, D_2, D_3, \ \ldots, D_M\}$ is a cluster of M documents to be summarised, we combine all sentences of the document collection to obtain a flattened cluster, $C = \{S_1, S_2, S_3, S_4, S_4, S_4, \ \ldots, S_N\}$, where N is the total number of cluster sentences. Next, a filtering process is applied to C in order to sieve cluster sentences by discarding all highly similar sentences to the current one. Fig. 4 describes the cluster merging process. For better readability, the figure indicates outward arrows for S_1 only, but the same logic applies to the rest of the sentences. By this merging, we remove $(N - K)$ sentences where $N \geq K$.

5.2. Computing SRL-ESA based semantic similarity

The fundamental building block of our summarization system is the determination of the role-based semantic similarity for intra-sentence similarity graphs and redundancy avoidance. Developing more effective sentence similarity measures are now believed to hold the future potential of extractive text summarization (Mehta & Majumder, 2018). To calculate the semantic similarity, we first pre-processed documents by merging each collection of related documents (multi-document summarisation only) and then segmented both single and multi-documents into sentences. Next, we constructed the semantic representation of each sentence by parsing it with semantic role labelling software. This semantic parsing aims at discovering semantic frames and associated arguments for each document sentence. The semantically parsed sentences are then formatted to a custom template for subsequent processing.

For exemplification, consider Example 2 of highly semantically related sentences.

Example 2.

S_1: FIFA was accused of corruption.
S_2: FIFA was officially investigated for corruption.

Applying semantic parsing identifies the predicate verbs of each sentence. In this case, each sentence has a single predicate verb, *accuse* for sentence 1 and *investigate* for sentence 2, and hence one primary semantic frame each. The role set of each predicate is classified according to the semantic roles they sit with respect to the verb. With this respect, three arguments, namely, A1 (direct

Table 3
Tokenised Example 2 sentences with their predicates and semantic role tags.

S_1 predicates and semantic arguments			**S_2 predicates and semantic arguments**		
Terms	Predicates	Role tags	Terms	Predicates	Role tags
FIFA	–	A1	FIFA	–	A1
was	–	0	was	–	0
accused	*accused*	V	officially	–	AM-MNR
of	–	B-A2	investigated	*investigated*	S-V
corruption	–	E-A2	for	–	B-A2
			corruption	–	E-A2

Table 4
Role-Term(s): common semantic roles and their corresponding term vectors.

Role (Arg.) label	**S_1 argument terms (WV_{i1})**	**S_2 argument terms (WV_{i2})**
V	Accuse	Investigate
A1	FIFA	FIFA
A2	corruption	corruption

object), A2 (indirect object) and AM-MNR (manner) are identified in both sentences. Table 3 shows a breakdown of both sentences in Example 2 into semantic frames indicating the semantic role that each token fills in the predicate.

Formally, let S_1 and S_2 be two sentences consisting of semantic frames f_1 and f_2 respectively. Let $R_1 = \{r_1, r_2, ..., r_k\}$ and $R_2 = \{r_1, r_2, ..., r_l\}$ be the semantic role sets associated with f_1 and f_2 where k, and l are the numbers of arguments in the semantic frames. From the two role sets of the semantic frames, we select the common roles, $R_c = r_1, r_2, ..., r_m$, co-occurring in both sentences. All other unshared semantic roles are discarded from the calculation of the semantic similarity. This is because of the intuition that an accurate similarity can be captured by comparing the semantic arguments corresponding to matching semantic roles. Having identified all shared semantic roles, the next step of our similarity computation involves building a Role-Terms Table for each sentence. The Role-Terms Table is a table that lists all shared semantic roles along with their related term vectors.

For instance, if we assume that $TV = \{WV_{1i}, WV_{2i}, ..., WV_{mi}\}$ are term vectors related to the semantic roles $r_1, r_2, ..., r_m$ of sentence i, the Role-Terms Table for Example 2 can be constructed as in Table 4, a better organization of the data in Table 3. The table shows argument terms of the shared roles for the example sentences after normalizing tokens, removing the noise (stop) words, and leaving semantic content words. Since there are few words in the example pair, we created a single Role-Terms Table for both sentences.

Once Role-Terms are constructed, the next step of our SRL-ESA based semantic similarity calculation is to translate the argument terms to their corresponding Wikipedia concepts. This is aided by a pre-built inverted index file containing a mapping of English content words to a weighted vector of hosting Wikipedia concepts as described in the preceding section. Continuing from our previous discussion, we interpret the Role-Terms Table to a table of concept vectors where each concept vector replaces argument terms filling the same semantic role. If WV_{ij} represents the argument term(s) of role i from sentence j, it translates to CV_{ij}, the weighed vector of Wikipedia concepts corresponding to WV_{ij}. For illustration purpose, Table 5 shows the first 5 Wikipedia concepts corresponding to the argument term *corruption* along with their unique Wikipedia ID numbers and TF-IDF weights.

Table 5 demonstrates the interpretation of the argument terms to corresponding Wikipedia concept vectors. Next, the actual semantic similarity between the two sentences is computed using these representative natural concepts. If $r_1, ..., r_m$ denote the shared semantic roles between two sentences where m is the number of the common roles, we use the Wikipedia concept vectors translated from the argument terms filling in these semantic roles. More formally, let $\{CV_{1k}, ..., CV_{ik}\}$ and $\{CV_{1l}, ..., CV_{il}\}$ be the concept vectors interpreted from the argument terms of the common roles between sentences k and l. The semantic similarity between sentences k and l is calculated as the average role similarities (RSim) obtained from the corresponding shared role sets. This is defined in Eq. (3) where i denotes the shared roles.

$$Sim_{srl-esa}(S_k, S_l) = \frac{1}{m}\sum_{i=1}^{m} RSim(CV_{ki}, CV_{li}) \tag{3}$$

The $RSim(CV_{ki}, CV_{li})$ is computed using individual concepts representing the original argument terms as formulated in (4). In Eq. (4), wc_{jk} represents the *tf*idf* weight of term j with respect to its corresponding concept from argument role i of sentence k while wc_{jl} is the *tf*idf* weight of term j with respect to its corresponding concept from argument role i of sentence l.

$$RSim(CV_{ki}, CV_{li}) = \frac{\sum_{j=1} wc_{jk} * wc_{jl}}{\sqrt{\sum_{j=1} wc_{jk}^2}\sqrt{\sum_{j=1} wc_{jl}^2}} \tag{4}$$

Fig. 5 summarizes the SRL-ESA based procedure for calculating the semantic similarity between two short texts ST_1 and ST_2. The figure demonstrates four procedural stages with the assumption of three shared roles.

1. The first step applies the semantic parsing by using semantic role labelling (SRL). The input to this stage is a short text and the output is a semantically tagged/parsed text.

Table 5
Top 5 Wikipedia concepts of the argument term: *Corruption*.

Wikipedia ID#	Concepts	TF*IDF weight
20055663	Prevention of Corruption Act	0.5399063230
2110801	Corruption (linguistics)	0.5140590668
25239439	Corruption in the United States	0.4959531128
3174020	Corruption Perceptions Index	0.45036080479
66241	Transparency International	0.4280707538

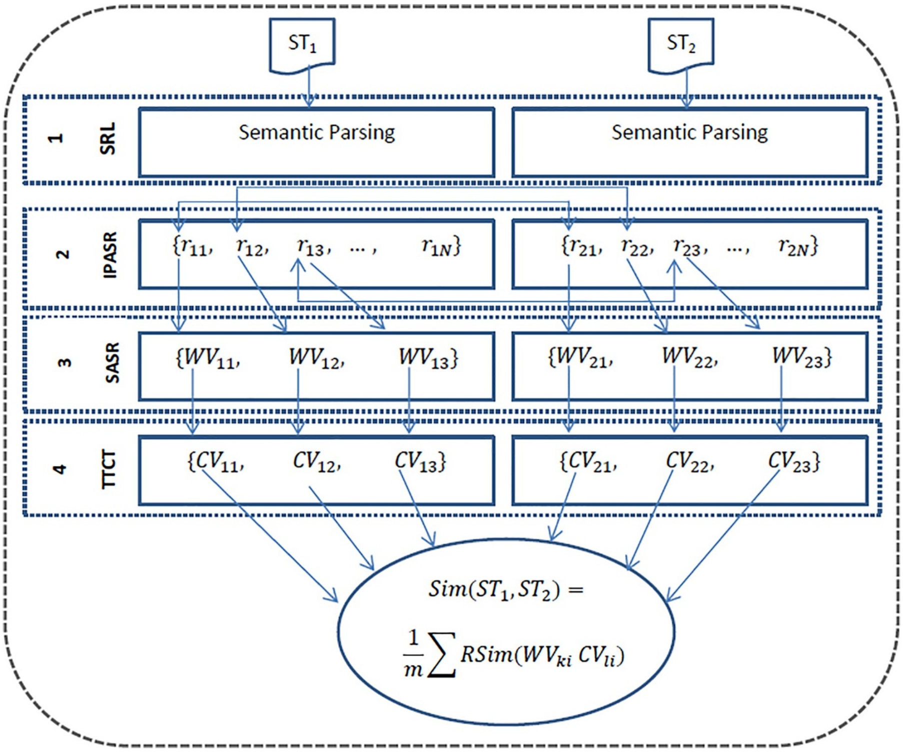

ST_1
ST_2
1 SRL
Semantic Parsing
Semantic Parsing
2 IPASR
$\{r_{11}, r_{12}, r_{13}, \dots, r_{1N}\}$
$\{r_{21}, r_{22}, r_{23}, \dots, r_{2N}\}$
3 SASR
$\{WV_{11}, WV_{12}, WV_{13}\}$
$\{WV_{21}, WV_{22}, WV_{23}\}$
4 TTCT
$\{CV_{11}, CV_{12}, CV_{13}\}$
$\{CV_{21}, CV_{22}, CV_{23}\}$
$Sim(ST_1, ST_2) = \frac{1}{m}\sum RSim(WV_{ki}\ CV_{li})$

Fig. 5. SRL-ESA based semantic similarity computation for short texts.

2. Secondly, the predicate verbs for the text are detected together with their semantic role sets. Therefore, this stage is called Identification of Predicates and Associated Semantic Roles, shortly abbreviated as IPASR.
3. Thirdly, our process recognises that all semantic roles are not shared in a typical short text and selects the arguments of common semantic roles in the third stage. This is referred to as Selecting Arguments of Shared Roles (SASR).
4. The final stage translates all grouped argument terms to their corresponding weighted Wikipedia concepts before carrying out the actual similarity calculation. This stage is known as Terms to Concepts Translation, or TTCT.

5.3. Semantic graph representation of documents

In the next step of the proposed summarization approach, every document is represented as a weighted undirected graph where the sentence concepts form the nodes (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: α: $V \to \mathbb{R}_+$ and β: $E \to \mathbb{R}_+$ are functions defining the vertex rankings and 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_i) of the semantic argument terms. Fig. 6 shows the semantic argument representation (A) and the sentence level similarity (B) graphs used for similarity computation and sentence ranking,

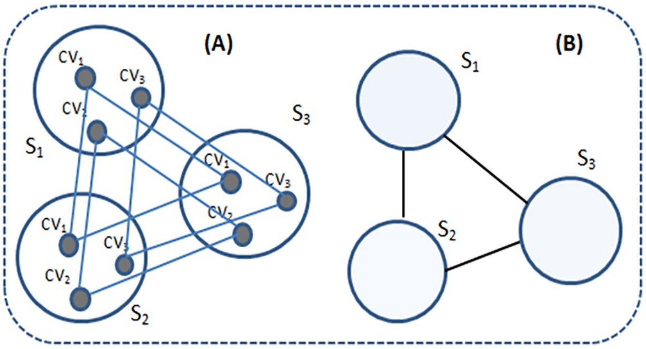


Fig. 6. Semantic argument level (A) and sentence level (B) document similarity graphs.

respectively. The graph vertices are ranked using PageRank algorithm (Brin & Page, 2012). The edge-weights are formulated as per Eq. (5). For single document summarization, the weights are measured in a slightly different way than for 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 semantic similarity with the document title indicates an additional importance of that given sentence in the document (see the document example in Fig. 7 and Tables 7 and 8).

$$EdgeWeight(\beta) = \begin{cases} Sim_{srl-esa}(S_1, S_2) + TSim(S_1, S_2, T) & for\ SDS \\ Sim_{srl-esa}(S_1, S_2) & for\ MDS \end{cases} \tag{5}$$

where S_i denotes sentence *i*, *T* is the document title, SDS (resp. MDS) represents single document summarization (resp. multi-document summarization), $Sim_{srl-esa}(S_1, S_2)$ is the SRL-ESA based similarity measure formulated in Eq. (3) and $TSim(S_1, S_2, T)$ is the tile similarity which is given in (6).

$$TSim(S_1, S_2, T) = 0.5*(Sim_{esa}(S_1, T) + Sim_{esa}(S_2, T)) \tag{6}$$

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, sentences without predicate verbs are not included in the graph representation. This is because the SRL based semantic analysis cannot be applied to such sentences lacking semantic frames.

5.4. Iterative sentence ranking and summary extraction

We applied PageRank algorithm to the document similarity graphs to rank and identify the most important sentences to be extracted as a summary. The PageRank for page p_i, $PR(p_i)$, is formulated as in Eq. (7) where $In(p_i)$ and $Out(p_j)$ are the total numbers of incoming and outgoing links for pages p_i and p_j respectively, N represents the total number of pages, and λ is the probability that an Internet surfer will continue navigating to other pages randomly, known as a damping factor. The recommended value for λ is 0.85 but can be set to any number between 0 and 1.

$$PR(p_i) = \frac{1-\lambda}{N} + \lambda * \sum_{p_j \in In(p_i)} \frac{PR(p_j)}{Out(p_j)} \tag{7}$$

In the summarization context, we rank document sentences instead of web pages; hence sentences play the roles of webpages. For a document graph, intra-sentence semantic similarities take the place of incoming and outgoing links in the computation of sentence ranks. The rank of each sentence indicates its salience which depends on the number and the strength of semantic links connecting each sentence to the rest of the document sentences. In other words, sentences with strong connections (high semantic similarities) are more likely to be candidates for summary inclusion than those with a lot of weak connections (low similarities). To exemplify our reasoning, a short document of 10 sentences taken from the DUC 2002 dataset is represented in Fig. 7. To apply the proposed approach, the document is first pre-processed and parsed with SRL to allow the extraction of predicate verbs (semantic frames) and associated semantic role-sets. Fig. 8 shows the sentence similarity graph of the document. Note that, for the purpose of legibility, self-links and intra-sentence links with similarity scores below ***0.05*** are omitted from Fig. 8. However, Table 6 lists the SRL-ESA based intra-sentence similarity scores of the entire document. In addition, the numbers in Table 7 represent the title-sentence similarities whose averages are combined with intra-sentence similarities to form the edge weights in accord with Eqs. (5) and (6). Finally, Table 8 shows final sentence ranks after running the PageRank algorithm for 20 iterations, which is where the ranking algorithm empirically reached its steady state. The impact of the title-sentence similarity scores is evident on the sentence rankings where sentences with title similarities, namely 1, 2, 9 (see Table 7) are highest ranked as shown in Table 8, alluding to the importance of the sentences semantically connected with the document title.

From the given sentence ranking scores in Table 8, the document sentences are ranked according to their importance as in the rank order row of the table with the most and least salient sentences being the first and the seventh respectively. For the purpose of

Table 6
Intra-sentence similarities of the example document in Fig. 7 using the SRL-ESA based semantic similarity measure.

Sent. No.	1	2	3	4	5	6	7	8	9	10
1	1.0									
2	0.1436	1.0								
3	0.0690	0.0458	1.0							
4	0.0472	0.1320	0.0234	1.0						
5	0.0477	0.0210	0.0246	0.0217	1.0					
6	0.1123	0.0789	0.1054	0.26	0.0044	1.0				
7	0.0023	0.0012	0.0137	0.0059	0.0033	0.0517	1.0			
8	0.2399	0.0685	0.0654	0.2733	0.0683	0.1279	0.0071	1.0		
9	0.4692	0.2514	0.0407	0.3291	0.0212	0.0081	0.0	0.3513	1.0	
10	0.00025	0.0002	0.1297	0.0001	0.0015	0.0	0.0	0.0015	0.0	1.0

Table 7
Sentence-title similarity scores of the example document in Fig. 7.

Sentence no.	1	2	3	4	5	6	7	8	9	10
Title similarity score	0.3333	0.13094	0.0	0.0	0.0	0.0	0.0	0.0	0.3448	0.0

Table 8
Sentence-ranks after 20 iterations of the example document in Fig. 7.

Sentence no.	1	2	3	4	5	6	7	8	9	10
Sentence ranks	1.446	1.095	0.912	0.968	0.862	0.938	0.85	0.979	1.057	0.894
Rank order	1	2	7	5	9	6	10	4	3	8

Document ID: AP000825-0099, Cluster ID: d070f, Dataset: DUC 2002
Title: Report - Honecker Unlikely To Go to Trial in East Germany

S_1: Ousted East German leader Erich Honecker will not stand trial in East Germany as long as the formerly Communist country exists, a West German newspaper reported.

S_2: The Hamburg-based Bild am Sonntag said Saturday that it would report in its Sunday editions that Honecker could be prosecuted in a united Germany, however, for violation of property laws.

S_3: Bild quoted Guenter Seidel, an East German prosecutor, as saying that Honecker had used 42 million for stocking a private housing estate for leaders of the former Communist government.

S_4: However, Seidel said that the investigation was not far enough along to determine whether charges could be filed against Honecker before East Germany merges with West Germany on Oct.

S_5: Negotiators are still working out the merger of the two German legal systems.

S_6: Honecker, 78, was ousted as East Germany's leader on Oct. 18, paving the way for the country's first freely elected government in March.

S_7: Honecker is in poor health and remains confined to a Soviet military hospital in Beelitz outside East Berlin.

S_8: He is under investigation on allegations of abuse of power, corruption, harboring terrorists and issuing shoot-to-kill orders to prevent East Germans from escaping to West Germany when he served as the country's leader.

S_9: Bild said that Erich Mielke, the ex-head of East Germany's former secret police, was also unlikely to go to court in East Germany.

S_{10}: I am at the end. I am a dead man, Bild quoted Mielke, 82, as saying at his last interrogation.

Fig. 7. An example document to be summarized.

summary generation, the highest ranked sentences of not more than the required summary length (100 words), as set by the DUC 2002 guidelines, are selected as a summary. With this restriction in action, the extracted summary as given in Fig. 9 comprises of sentences 1, 2, 9, and part of sentence 8.

6. Experiments

6.1. Evaluation

For the purpose of testing and validation, we used 21 clusters of 160 documents semi-randomly selected mainly from the first half of the DUC 2002 corpus. Table 9 provides some statistics of the evaluation dataset. Each document or cluster in the DUC 2002 dataset comes with a model summary of various lengths ranging from 10 $\sim$ 400 words, which are either created or extracted by human experts to serve as a reference summary. The overall experimental design of the proposed summarizer is shown in Fig. 3 and detailed

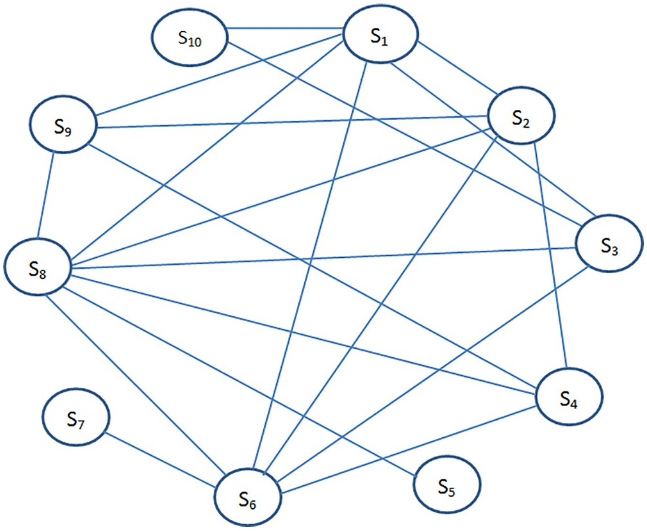
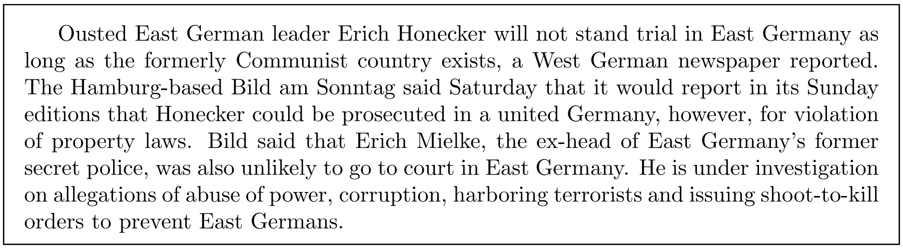


Fig. 8. Sentence similarity graph of the example document in Fig. 7; for clarity, the graph only shows links with weights ≥ 0.05, while the full link values are given in Table 6.



Ousted East German leader Erich Honecker will not stand trial in East Germany as long as the formerly Communist country exists, a West German newspaper reported. The Hamburg-based Bild am Sonntag said Saturday that it would report in its Sunday editions that Honecker could be prosecuted in a united Germany, however, for violation of property laws. Bild said that Erich Mielke, the ex-head of East Germany's former secret police, was also unlikely to go to court in East Germany. He is under investigation on allegations of abuse of power, corruption, harboring terrorists and issuing shoot-to-kill orders to prevent East Germans.

Fig. 9. A 100-word summary extracted from the example document in Fig. 7.

Table 9
Statistical description of the evaluation dataset.

Cluster	Document #	Sentence #	Word #	Task
D061j	6	238	3933	SDS & MDS
D062j	5	158	2869	SDS & MDS
D064j	7	254	4398	SDS & MDS
D065j	8	371	5890	SDS & MDS
D066j	7	250	4127	SDS & MDS
D067f	6	168	2984	SDS & MDS
D068f	5	182	2791	SDS & MDS
D070f	11	250	3563	SDS & MDS
D071f	6	204	2380	SDS & MDS
D072f	13	483	8343	SDS & MDS
D074b	6	316	4057	SDS & MDS
D075b	10	328	6261	SDS & MDS
D076b	10	421	6765	SDS & MDS
D077b	10	429	6842	SDS & MDS
D079a	9	398	6873	SDS & MDS
D080a	11	653	10868	SDS & MDS
D081a	12	403	6604	SDS & MDS
D083a	6	276	4622	SDS & MDS
D108g	10	296	4081	SDS & MDS
D109h	10	275	4262	SDS & MDS
D113h	5	183	2833	SDS & MDS

in Section 5. As for the quantitative evaluation of the system against baselines and related works, we employed the Recall Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which is the most widely used official evaluation tool in text summarization. As given in Eq. (8), the ROUGE determines the quality of a system summary by comparing its text to an ideal human summary (aka as a model/reference summary) and computing a group of ROUGE measures including ROUGE-N, ROUGE-SU, and ROUGE-L.

$$ROUGE - N = \frac{\sum_{S \in RefSumm,} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in RefSumm,} \sum_{gram_n \in S} Count\,(gram_n)} \tag{8}$$

Where N is the length of the n-gram ($gram_n$), $Count(gram_n)$ is the number of n-grams ($gram_n$) in the reference summary while $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the system summary and the collection of reference summaries (RefSumm). A $gram_n$ refers to a sequence of n words, for instance two-word is called a bigram.

6.2. Results and discussion

To extract a document summary S, we made use of the semantic graph interconnectivity among document sentences to calculate a quality ranking for each sentence. To this end, PageRank algorithm is iteratively run on the document similarity graphs (see Section 5.4) until it converges. All sentences are ranked equally at the beginning of the algorithm, which is run recursively on document similarity graphs until it reaches a steady state. Then, each sentence is ranked depending on the number of other connected sentences and the strength of the similarity between it and the rest of the document or cluster sentences. Sentences with high semantic similarity and linked with many other document sentences are favoured and ranked higher. These are then sorted according to their ranks and selected as a summary.

Following NIST's [4] dataset construction guidelines, the lengths of extracted summaries are 100 and 200 words for SDS and MDS respectively. Tables 10 and 11 show the quality of the system summaries produced for SDS and MDS in terms of the average ROUGE recall scores. The choice of the measures is made on the basis of the findings in Lin (2004), where researchers reported that the measures used for Qf-MDS are the ones that work well for topic-focussed MDS and that the measures, ROUGE-N (N = 1, 2), ROUGE-L, and ROUGE-SU4 effectively reflect the effectiveness of generic SDS systems. The numbers in the square brackets following the ROUGE scores are minimum and maximum recorded ROUGE recall values in the format [min-max]

6.2.1. Generalization and data size effect

To draw a generalization from the obtained results, we investigated the impact of the data size on the performance of the summarisers. Tables (12) and (13) illustrate how changing data size, in terms of the number of documents for SDS (Table 12) and the number of clusters for the MDS (Table 13), affects the summariser's performance. Visualizations of the same results are also shown in Fig. 10. Interestingly, we found that system results when tested on varying data sizes remain consistent on average, particularly for the SDS task. However, experimental results seem to be less consistent for the MDS as compared to the SDS task. This is shown in Table 13 where the scores of the ROUGE measures have larger deviations compared to the corresponding scores of the SDS task in Table 12. Possibly, this is because of the large document sizes, in terms of the number of sentences after merging multi-document clusters. It could also be due to the high compression rate needed to summarize long merged cluster documents. Succinctly, this set of results indicate that the variation of the evaluation data size has little influence on the quality of the summaries, particularly for the single-document summarization. Therefore, we may deduce that the proposed SRL-ESA graph-based SDS and MDS system is scalable, which leads us to generalize that the evaluation can represent a dataset of any size.

A very commonly used statistical technique for generalizations though is the concept of confidence intervals (CI). It is the range of values that is thought to include the true representative value, or the mean, of the entire results. In our case, that figure is the average ROUGE score. Luckily, for our results, this generalization has been implemented in the ROUGE measure where a bootstrap resampling technique is applied to generalize evaluation results (Lin, 2004). Specifically, it uses a 95% confidence interval, which indicates the range within which any result in the evaluation is true 95% of the time for the entire data.

6.2.2. Comparison with benchmark methods and related works

In the previous section, we have shown how changing the data size has little impact on the performance of the proposed summarizer, which allowed us to generalize system results. In this section, we compare our results with benchmark methods including Microsoft Word Summariser and related state-of-the-art systems. The selection of the aforementioned comparators was based on their relatedness to the current work in terms of the implementation and the evaluation dataset. In the following paragraphs, we provide a brief introduction to each of the used comparators.

- **Microsoft Word summariser** is a summarisation tool embedded in the Microsoft Word Application. This summarizer uses term frequencies to calculate the relative importance of each sentence in a document. It is used in some related studies (Abuobieda et al., 2013; Binwahlan et al., 2010; Suanmali et al., 2010) as a benchmark method for automatic summarisation systems.
- **System 19** is the best performing summarization system at the relevant competition in the Document Understanding Conference (DUC) (Harabagiu & Lacatusu, 2002). The summarizer, which was implemented for the evaluation of the DUC 2002 dataset, can be be used to produce summaries from both single and multiple documents.
- Binwahlan et al. (2010) is a hybrid model for automatic single-document summarization combining diversity, swarm and fuzzy methods to achieve three criteria; avoiding redundancy in the generated summary, identification of the most appropriate text

[4] National Institute of Standards and Technology.

Table 10
The overall results of the SRL-ESA graph-based single document summarisation (SDS): average recall of the four selected ROUGE measures at 95% confidence interval.

Measure	Recall	Precision	F-measure
ROUGE-1	0.504 [0.228–0.790]	0.431 [0.212–0.665]	0.462 [0.232–0.676]
ROUGE-2	0.235 [0.029–0.537]	0.201 [0.023–0.510]	0.216 [0.026–0.518]
ROUGE-L	0.335 [0.132–0.592]	0.286 [0.101–0.560]	0.307 [0.121–0.572]
ROUGE-SU4	0.254 [0.062–0.534]	0.216 [0.061–0.487]	0.232 [0.063–0.496]

Table 11
The overall results of the SRL-ESA graph-based multi-document summarisation (MDS): average recall of the three selected ROUGE measures at 95% confidence interval.

Measure	ROUGE-1	ROUGE-2	ROUGE-SU4
Recall	0.474 [0.336–0.642]	0.212 [0.068–0.380]	0.246 [0.106–0.413]
Precision	0.427 [0.318–0.529]	0.190 [0.064–0.323]	0.220 [0.100–0.336]
F-measure	0.449 [0.327–0.577]	0.201 [0.066–0.341]	0.232 [0.103–0.371]

Table 12
Impact of data size (number of documents - NoD) for Single Document Summarization (SDS) - all figures are rounded to 3 significant figures.

NoD	ROUGE-1	ROUGE-2	ROUGE-L
10	0.492	0.206	0.290
20	0.475	0.211	0.300
30	0.501	0.253	0.337
40	0.503	0.247	0.334
50	0.504	0.245	0.337
60	0.495	0.228	0.321
70	0.469	0.229	0.321
80	0.492	0.225	0.319
90	0.488	0.218	0.315
100	0.494	0.226	0.324
110	0.497	0.229	0.328
120	0.504	0.235	0.335

Table 13
Impact of data size (number of clusters - NoC) for Multi Document Summarization (MDS) - all figures are rounded to 3 significant figures.

NoC	ROUGE-1	ROUGE-2	ROUGE-SU4
1	0.471	0.197	0.229
2	0.336	0.068	0.106
3	0.417	0.137	0.183
4	0.642	0.380	0.413
5	0.587	0.359	0.370
6	0.513	0.248	0.274
7	0.430	0.105	0.158
8	0.502	0.231	0.259
9	0.469	0.282	0.305
10	0.378	0.117	0.160

features for sentence scoring and the optimization & adjustment of feature weights.

- Wan (2010) is system which combines SDS and MDS tasks by examining the mutual influences between them. It proposes a unified graph model called CoRank which establishes word-sentence relationship. The proposed approach combines local saliency for the identification of sentence importance in a particular document with global saliency used to indicate the importance of a sentence in a cluster.
- Abuobieda et al. (2013) is a summarization approach that combines opposition-based machine learning and differential evolution. The opposition-based learning is used to optimize the performance of Differential Evolution (DE), which itself is used to adjust the algorithm's initial population instead of relying on a random number generator.
- Alguliyev et al. (2013) is an optimization-based multi-document summarization model. It uses sentence-to-document collection, summary-to-document collection and sentence-to-sentence relations to select salient sentences and reduce redundancy in the

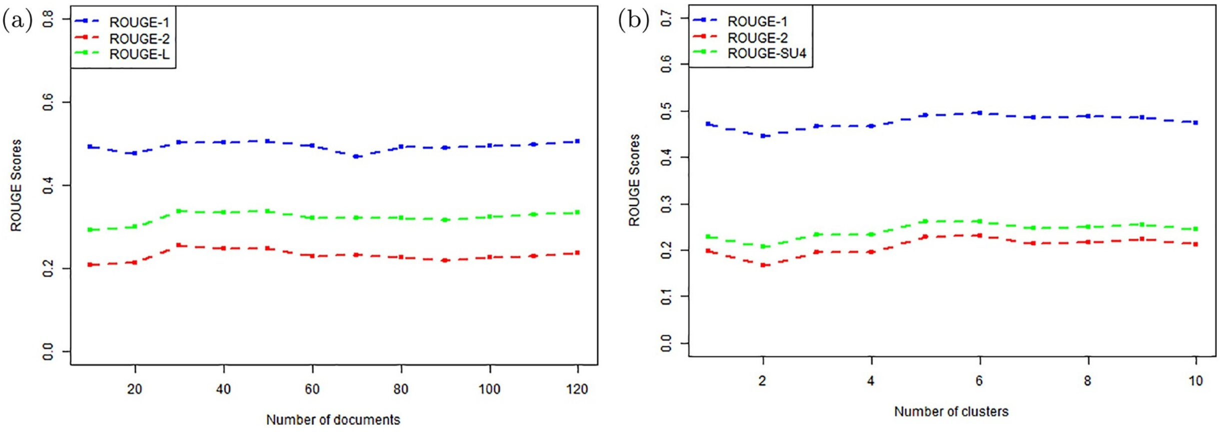


Fig. 10. Impact of data size on the SRL-ESA graph-based SDS (a) and MDS (b) tasks.

summary.

- Kumar et al. (2014) is a topic-focused multi-document summarization strategy based on cross document relations and fuzzy reasoning. The approach follows three phases; extracting news components in the documents using WordNet thresasure, named entity recognition and Gazateer lists, establishing cross document relations to identify relevant sentences, and finally the application of fuzzy reasoning to assign final sentence scores.
- Sankarasubramaniam et al. (2014) is a graph-based summarization system leveraged with Wikipedia concepts. The study maps document sentences to Wikipedia concepts for the construction of bipartite sentence-concept graphs. The summarizer then ranks sentences based on the ranking of corresponding concept nodes. It is noteworthy that this is one of the closest studies to our work in terms of the implementation, e.g., the use of concept graphs and Wikpedia as external knowledge.
- Oliveira et al. (2016) is a generic single and multi document summarizer based on eighteen of the most widely used sentence scoring techniques including TextRank, TF-IDF, graph-based similarity, named-entities, sentence centrality and word co-occurrence. The scoring methods are used to compute the sentence importance in a document/cluster. This comparative study found that the strategy of combining features can lead to improved results.
- Ren et al. (2016) is a redundancy-aware summarization system which considers the importance of the sentences and the redundancy in the summary simultaneously instead of modelling them as two separate processes. Particularly, this approach first evaluates the importance of each sentence and then selects sentences to generate a summary based on both the importance scores and redundancy among sentences.
- Sun and Zhuge (2018) is a SDS approach which extracts semantic link network from documents using graph representations of different language granularities (words, sentences, paragraphs and sections) as nodes and semantic links between the nodes. In this method, *is-part-of, similar-to* and *co-occurrence* links are built among various nodes to form Semantic Link Network for modelling the basic semantic structure of a document.
- Alguliyev et al. (2018) is a two-stage sentence selection summarization model based on clustering and optimization techniques. The first stage discovers all topics in a text by clustering the sentences set using k-means method. In the second stage, optimization is employed to model the selection of salient sentences from document clusters.

Tables 14 and 15 include a comparison of our results against baselines and other related state-of-the-art works for SDS and MDS tasks respectively. The numbers in the parenthesis following the scores of the ROUGE measures represent the ranking of each system in the list. As shown, the proposed system ranks the top in the ROUGE-1 and ROUGE-2 measures for the SDS task, while ranking the

Table 14
Comparison of our results with benchmark methods and related works for Single Document Summarization (SDS) - all figures are rounded to 3 significant figures.

Summarizer	ROUGE-1	ROUGE-2	ROUGE-L
This system	**0.504 (1)**	**0.235 (1)**	0.335 (3)
MS Word	0.471 (5)	0.212 (8)	0.310 (5)
System 19	0.459 (8)	0.233 (2)	0.313 (4)
Wan (2010)	0.485 (3)	0.215 (7)	—
Binwahlan et al. (2010)	0.436 (10)	0.197 (10)	0.401 (2)
Abuobieda et al. (2013)	0.445 (9)	0.224 (5)	**0.407 (1)**
Sankarasubramaniam et al. (2014)	0.460 (6)	0.230 (4)	—
Oliveira et al. (2016)	0.477 (4)	0.223 (6)	—
Sun and Zhuge (2018)	0.460 (7)	0.207 (9)	—
Alguliyev et al. (2018)	0.491 (2)	0.231 (3)	—

Table 15
Comparison of our results with benchmark methods and related works for Multi-Document Summarization (SDS) - all figures are rounded to 3 significant figures.

Summarizer	ROUGE-1	ROUGE-2	ROUGE-SU4
This system	0.474 (2)	0.212 (3)	0.246 (2)
MS Word	0.451 (4)	0.198 (4)	0.242 (4)
System 19	0.467 (3)	0.213 (2)	0.245 (3)
Wan (2010)	0.383 (5)	0.079 (7)	—
Alguliyev et al. (2013)	**0.499 (1)**	**0.255 (1)**	**0.286 (1)**
Kumar et al. (2014)	0.332 (8)	0.128 (5)	0.100 (5)
Oliveira et al. (2016)	0.358 (7)	0.078 (8)	—
Ren et al. (2016)	0.378 (6)	0.0961 (6)	—

second place in the ROUGE-1 and ROUGE-SU4 measures for the MDS task. It is also clear that the proposed system is placed third based on the other two measures, namely the ROUGE-L (SDS) and the ROUGE-2 (MDS). This indicates the competency of the proposed SRL-ESA Wikpedia graph-based summarisation where it excelled in terms of ROUGE scores as compared to the benchmark and related state-of-the-art summarisation methods. However, it goes without saying that the proposed approach underperforms the work of Alguliyev et al. (2013) in the MDS task, as shown in Table 15.

Overall, the system's experimental results underline the advantages of the proposed summarization approach and the use of crowd-sourced knowledge with sentence-level semantic parsing for single and multi document summarisation tasks. In other words, the findings highlight the importance of using semantic parsing and argument terms' concepts to compute intra-sentence similarities. It also opens the door to further exploration of the applicability of our SRL-ESA Wikipedia based similarity measure to other similarity underpinned NLP tasks including paraphrase identification, plagiarism detection, conversational agents, text classification, among others. In addition, establishing semantic links under sentence level (i.e. the multi-node representation of sentence vertices) could be used for other NLP tasks such as concept extraction where ranking argument concepts is more appropriate than ranking sentences. Similarly, the proposed summarization framework can be applied to other languages included in the Wikipedia database provided the availability of equivalent semantic analysis techniques for that language.

7. Conclusion and future work

In this paper we proposed a text summarization approach encompassing both single-document and multi-document summarization using semantic role labeling and Wikipedia-based explicit semantic analysis. The SRL is used for the semantic representation of document sentences while the ESA algorithm facilitates the interpretation of semantically parsed sentences to indexed Wikipedia concepts. Semantic roles are paired if they fill the same semantic position in a sentence. Argument texts pertaining to the shared semantic roles are then projected to a vector of corresponding Wikipedia concepts where the intra-sentence semantic relatedness is computed from such concept vectors. A graph-based SDS & MDS is built on the basis of the developed SRL-Wikipedia based similarity measure. The paper also presented an experimental evaluation of the proposed methodology on a standard publicly available dataset from the relevant DUC conference. The obtained results revealed considerable performance improvements in the summary quality illustrating the power of the role-based semantic representation and its mapping onto a human generated natural concepts encoded in Wikipedia. The findings also suggest that the other NLP tasks underpinned by semantic similarity functions can also be enhanced with this approach.

As a future work, we intend to apply the SRL-ESA Wikipedia-based method to a number of other summarization tasks including opinion, product/service review and guided summarisation. Our immediate interest is on the latter as guided summarization involves the retrieval of a summary response to an event described in a user question and is thought to be the best way of summarizing documents relating to topics of template-like categories, such as attacks, accidents and natural disasters, investigations, and health & safety. Such topics contain highly predictable facts such as *who did what when* and *where* and interestingly **SRL** can be the best tool for answering such event-based questions.

References

Abdi, A., Shamsuddin, S. M., & Aliguliyev, R. M. (2018). QMOS: query-based multi-documents opinion-oriented summarization. *Information Processing & Management, 54*(2), 318–338.

Abuobieda, A., Salim, N., Kumar, Y. J., & Osman, A. H. (2013). *Opposition differential evolution based method for text summarization. Intelligent information and database systems*. Springer Berlin Heidelberg487–496.

Aksoy, C., Bugdayci, A., Gur, T., Uysal, I., & Can, F. (2009). *Semantic argument frequency-based multi-document summarization. 24th International symposium on computer and information sciences (ISCIS)*. IEEE.

Alguliyev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications, 40*(5), 1675–1689.

Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2018). COSUM: Text summarization based on clustering and optimization. *Expert Systems,* e12340.

Allan, K. (2001). Natural language semantics.

Azadani, M. N., Ghadiri, N., & Davoodijam, E. (2018). Graph-based biomedical text summarization: An itemset mining and sentence clustering approach. *Journal of Biomedical Informatics, 84*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York: ACM press 463

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). *The berkeley framenet project. Proceedings of the 17th international conference on computational linguisticsVol. 1. Proceedings of the 17th international conference on computational linguistics* Association for Computational Linguistics86–90.

Binwahlan, M. S., Salim, N., & Suanmali, L. (2010). Fuzzy swarm diversity hybrid model for text summarization. *Information Processing & Management, 46*(5), 571–588.

Bollegala, D., Okazaki, N., & Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management, 46*(1), 89–109.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks, 56*(18), 3825–3833.

Canhasi, E., & Kononenko, I. (2011). *Semantic role frames graph-based multidocument sumarization. Proceedings of siKDD'11*.

Canhasi, E., & Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications, 41*(2), 535–543.

Carbonell, J., & Jade, G. (1998). *The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR), 12* Aug

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS), 29*(2), 8.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., & Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications, 41*(13), 5780–5787.

Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., & Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications, 40*(14), 5755–5764.

Fillmore, C. J., & Baker, C. (2010). *A frames approach to semantic analysis. The Oxford handbook of linguistic analysis*313–339.

Gabrilovich, E., & Markovitch, S. (2006). *Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. AAAIvol. 6. AAAI* 1301–1306.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research, 34*, 443–498.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review, 47*(1), 1–66.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics, 28*(3), 245–288.

Harabagiu, S. M., & Lacatusu, F. (2002). *Generating single and multi-document summaries with gistexter. Document understanding conferences*.

Jha, N., & Anas, M. (2018). Using frame semantics for classifying and summarizing application store reviews. *Empirical Software Engineering,* 1–34.

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using wikipedia. *Information Processing & Management, 51*(3), 215–234.

Khan, A., Salim, N., & Kumar, Y. J. (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing, 30*, 737–747.

Kumar, Y. J., Salim, N., Abuobieda, A., & Albaham, A. T. (2014). Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing, 21*, 265–279.

Lin, C.-Y. (2004). *ROUGE: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop*.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review, 37*(1), 1–41.

Mehta, P., & Majumder, P. (2018). Effective aggregation of various summarization techniques. *Information Processing & Management, 54*(2), 145–158.

Mihalcea, R., & Tarau, P. (2004). *TextRank: Bringing order into texts.* Association for Computational Linguistics.

Mohamed, M. (2016). *Automatic text summarisation using linguistic knowledge-based semantics*. University of Birmingham PhD Thesis.

Mohamed, M., & Oussalah, M. (2015). *Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources. Proceedings of the 9th IEEE international conference on big data science and engineering (IEEE bigdataSE-15)*.

Morales, L. P., Esteban, A. D., & Gervás, P. (2008). *Concept-graph based biomedical automatic summarization using ontologies. Proceedings of the 3rd textgraphs workshop on graph-based algorithms for natural language processing.* Association for Computational Linguistics.

Mosa, M. A., Anwar, A. S., & Hamouda, A. (2019). A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowledge-Based Systems, 163*, 518–532.

Nenkova, A., Maskey, S., & Liu, Y. (2011). *Automatic summarization. Proceedings of the 49th annual meeting of the association for computational linguistics, tutorial abstracts of ACL*. Association for Computational Linguistics.

Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., & Simske, S. J. (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications, 65*, 68–86.

Osman, A. H., et al. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing, 12*(5), 1493–1502.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics, 31*(1), 71–106.

Ren, P., Wei, F., Zhumin, C. H. E. N., Jun, M. A., & Zhou, M. (2016). *A redundancy-aware sentence regression framework for extractive summarization. Proceedings of COLING 2016, the 26th international conference on computational linguistics*33–43.

Saif, A., Omar, N., Ab Aziz, M. J., Zainodin, U. Z., & Salim, N. (2017). Semantic concept model using wikipedia semantic features. *Journal of Information Science, 44*(4), 526–551.

Sankarasubramaniam, Y., Ramanathan, K., & Ghosh, S. (2014). Text summarization using wikipedia. *Information Processing & Management, 50*(3), 443–461.

Suanmali, L., Salim, N., & Binwahlan, M. S. (2010). SRL-GSM: A hybrid approach based on semantic role labeling and general statistic method for text summarization. *Journal of Applied Sciences, 10*(3), 166–173.

Sun, X., & Zhuge, H. (2018). *Summarization of scientific paper through reinforcement ranking on semantic link network.* IEEE ACCESS.

Wan, X. (2010). *Towards a unified approach to simultaneous single-document and multi-document summarizations. Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics.

Wang, P., Hu, J., Zeng, H. J., & Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems, 19*(3), 265–281.

Wei, F., Li, W., Lu, Q., & He, Y. (2010). A document-sensitive graph model for multi-document summarization. *Knowledge and Information Systems, 22*(2), 245–259.

Yan, S., & Wan, X. (2014). SRRank: Leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22*(12), 2048–2058.

Zhou, Y., Guo, Z., Ren, P., & Yu, Y. (2010). *Applying wikipedia-based explicit semantic analysis for query-biased document summarization. International conference on intelligent computing.* Berlin, Heidelberg: Springer.

Zhuge, H. (2016). *Multi-dimensional summarization in cyber-physical society.* Morgan Kaufmann.