RMIT UNIVERSITY

Improving Single Document Summarization in a Multi-Document Environment

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Sharin Hazlin Huspi

B.Sc (Hons.), M.Sc

School of Science

College of Science, Engineering and Health

RMIT University

JUNE 2017

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Sharin Hazlin Huspi
27th June 2017.

Acknowledgments

First and foremost, I would like to thank Allah the Almighty, for all the blessings He has bestowed upon me so that I am able to finish this work.

Next, I would like to thank my supervisor, Prof. Mark Sanderson and Dr. Phil Vines, who gave me the support all along the way to get this thesis done. I am forever grateful with the motivation and encourgement in th four years under their guidance.

For families (especially my Mom and Dad), thanks for all your prayers and spirits. Heaps of love to my dearest husband, thank you for all the support and love. And to my two little kids, thanks for making me laugh when I really need it.

For my friends, I thank you for always being there for me.

Dedicated to my three i:
imie, izzul and izz.

Credits

Portions of the material in this thesis have previously appeared in the following publications:

- Yulianti, E., Huspi S.H., Sanderson, M. (2015). Tweet-based Sumamrization. Published in Journal of the Association of Information Science and Technology

This work was supported by the Malaysian Government and through the scholarship Kementerian Pendidikan Tinggi Malaysia and Universiti Teknologi Malaysia

The thesis was written in Microsoft Office 2016.

Note

Unless otherwise stated, all fractional results have been rounded to the displayed number of decimal figures.

Contents

5 Summary Evaluation: Relevant vs. Preference

6 Conclusion

Appendices

Bibliography

List of Figures

List of Tables

Abstract

Most automatic document summarization tools produce summaries from single or multiple document environments. Recent works have shown that there are possibilities to combine both systems: when summarising a single document, its related documents can be found. These documents might have similar knowledge and contain beneficial information in regard to the topic of the single document. Therefore, the summary produced will have sentences extracted from the local (single) document and make use of the additional knowledge from its surrounding (multi-) documents. This thesis will discuss the methodology and experiments to build a generic and extractive summary for a single document that includes information from its neighbourhood documents. We also examine the evaluation and configuration of such systems.

There are three contributions of our work. First, we explore the robustness of the Affinity Graph algorithm to generate a summary for a local document. This experiment focused on two main tasks: using different means to identify the related documents, and to summarize the local document by including the information from the related documents. We showed that our findings supported the previous work on document summarization using the Affinity Graph. However, contrary to past

suggestions that one configuration of settings was best, we found no particular settings gave better improvements over another. Second, we applied the Affinity Graph algorithm in a social media environment. Recent work in social media suggests that information from blogs and tweets contain parts of the web document that are considered interesting to the user. We assumed that this information could be used to select important sentences from the web document, and hypothesized that the information would improve the summary of a single document.

Third, we compare the summaries generated using the Affinity Graph algorithm in two types of evaluation. The first evaluation is by using ROUGE, a commonly used evaluation tools that measure the number of overlapping words between automated summaries and human-generated summaries. In the second evaluation, we studied the judgement of human users using a crowdsourcing platform. Here, we asked people to choose their judgement and explained their reasons to prefer one summary to another. The results from the ROUGE evaluation did not give significant results due to the small tweet-document dataset used in our experiments. However, our findings on the human judgement evaluation showed that the users are more likely to choose the summaries generated using the expanded tweets compared to summaries generated from the local documents only. We conclude the thesis with a study of the user comments, and discussion on the use of Affinity Graph to improve single document summarization. We also include the discussion of the lessons learnt from the user preference evaluation using crowdsourcing platform.

Chapter 1

Introduction

A summary is a text, which is a shorter version of the original document, where the main idea of the document is captured and presented in a simplified way. Therefore, a summary should be able to provide the important information, without having to read through the whole document. Humans are able to summarize documents themselves, but finding only the important and relevant information may take a lot of effort and time. Over the years, research in document summarization has grown due to readers needing help to reduce the amount of information that they encounter, even for normal daily tasks. Summaries have proven to be a significant first encounter for readers, such as in news headlines, reviews made for movies, books or song albums, and abstracts of scientific studies.

Since the 1950s, automatic summarization has been an active research topic. The challenge is to provide a summarization system that can screen and reduce the information to be read by humans. Most of the earlier work, focused on general, news-related documents where the aim of the summarization tasks was to have relevant and up-to-date information for the users. Nowadays, there are many different requirements and needs for document summarization. Thus, summarization tasks have become more user-specific and most of the works are focused on developing summarization tools that can satisfy a certain domain.

A summary can be produced by either extractive or abstractive methods. For many years, the

extractive methods using statistical approaches have proven to be easier to implement than the abstractive methods. This is because most of the statistical approaches focused on the use of word frequency to determine the most significant concepts within a document. Abstraction involves paraphrasing sections of the source document, which is challenging. Therefore, most of the document summarization systems available are extraction-based.

There are three factors affecting document summarization (Afantenos, Karkaletsis, & Stamatopoulos, 2005; Spärck Jones, 1993, 2007). First, the *input*: different numbers of documents will produce different types of summaries. Second, the *purpose* of a summary: it should be based on the audience and intended use of the summary (generic or user-oriented, general or domain specific). Third, the *output*: the summary can be in many different forms, such as paragraph or point form.

For the *input* factor, one should decide on the number of documents to be summarized, either single- or multi-document. Early work in summarization focused on generating summaries by using information contained only one document. Most such systems used extraction techniques to select the information regarded as most important and summarize each document individually. In recent years, there has been significant interest in multi-document summarization. It has received a lot of attention because groups of inter-related documents are more common. The process of selecting sentences from across the group is harder to realize.

In this thesis, the methods of improving automated summaries are investigated by exploring the possibilities of combining and extracting information from multiple documents for the purposes of improving single document summarization.

1.1 Motivation

In generic summarization, it is assumed that the reader would want to know the main content of the document, without reading the whole document itself. Thus, identifying the main topic(s) from multiple document to support single document summarization would be the main motivation to be discussed in this thesis, assuming that several topics can be found in a document.

Our approach is summarised in Figure 1.1. Given a document to be summarised, related documents must be selected from a collection (task 2 in Figure 1.1). Here, important topics would be comprehensively appeared in most of the sentences from the related documents; hence, it will be used to weights the sentences in the single document (task 3 in Figure 1.1). The highest weighted sentences are selected to produce an improved summary.

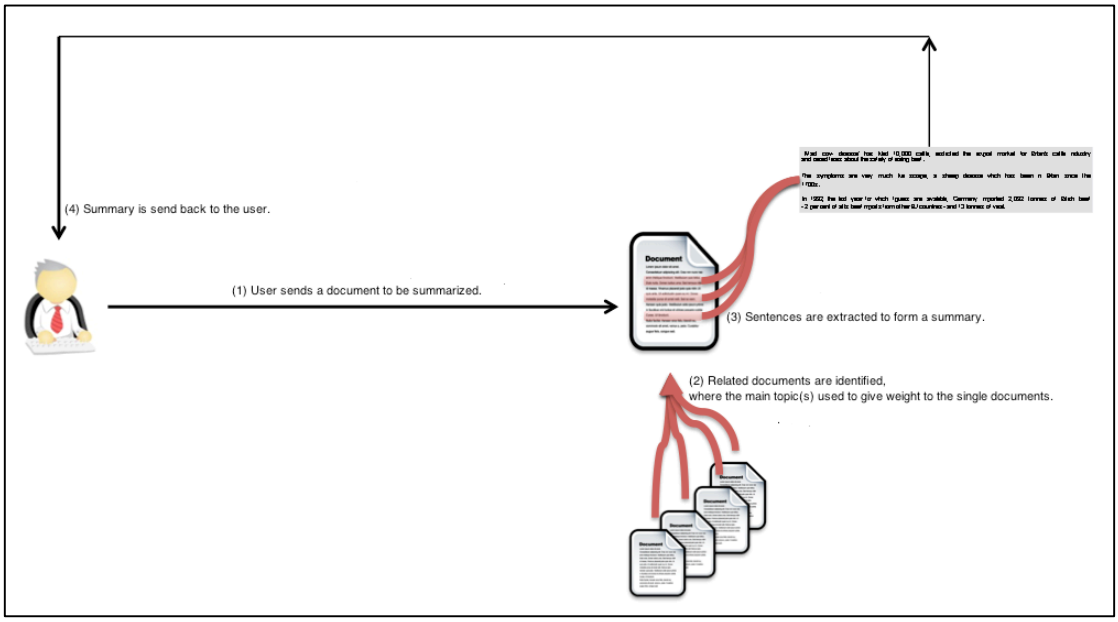


Figure 1.1: Improving single document summaries by identifying main topics from related documents

Another area of interest is document summarization evaluation. The standard way to evaluate a summary is by calculating its similarity to a reference summary. The reference summaries are created manually. Calculating similarity is challenging, as the definition of a good and fair gold-standard summary varies for different people and tasks. A 'good' auto-generated summary might be penalized (having low scores) because it might choose different words that did not appear in the gold standard summary; even though the selected word is from a similar topic.

In recent years, there has been an increasing interest in crowdsourcing, where a group of people (called participants/contributors/workers) were asked to do data labelling or judgment activities. Thus, it is expected that the work submitted to the crowdsourcing platform (mostly online)

would be faster and cost effective. Thus, we proposed that the use of crowdsourcing should be explored as an alternative to automated summarization evaluation.

1.2 Problem Statement

The main goal of the thesis is to answer the following problem:

How to extract sentences from one main text and use the information from its related documents to create a summary?

Goyal, Behera, & McGinnity (2013), Hu, Ji, Sun, Teng, & Zhang (2011) and Wan & Xiao (2010) discussed the potential of producing a summary that combined single document summarization with information from related documents. Their work applied a graph-based ranking algorithm, called an Affinity Graph (AG) to manage relationships between sentences extracted from both the documents to be summarized and a set of expanded documents (Zhang et al., 2005). In light of this past work, four research questions have been identified:

1. How effective are graph-based algorithms in improving single document summarization?
2. Can the same algorithm improve single document summarization using short text documents, in particular, Twitter messages?
3. Does a crowdsourced human judgement approach a better evaluation compared to the standard automated summary evaluation?
4. Will the new approach from (2) be preferred compared to a standard single document summarization?

We describe the work conducted in order to answer the four questions in more detail.

1.2.1 To Examine Different Means to Get Information from Surrounding Documents

The first research question will emphasize on the development of the summarization system. An Affinity Graph algorithm is used to explore the proposed idea, based on its success as presented in previous work.

Goyal et al. (2013), Hu et al. (2011) and Wan & Xiao (2010) examined a particular range of options in how to configure their summarization system. In this study, our interest is to explore the robustness of their methodology: can the approach be relied upon to yield improvement in a range of settings or does it only work in specific limited situations?

We investigated two possible mechanisms to improve single document summarization using additional knowledge from related documents. First, is to find the best way to get information from the surrounding documents. Similarity search, such as Cosine Similarity, Okapi BM25 and Language Modeling, are able to identify documents that are topically related to each other. Second, is to examine how much information from the related documents is needed to get the most knowledge from the collection. To investigate this, different numbers of related documents and variations in document length from the related documents will be considered.

In order to evaluate the summaries, we analysed the results in two different approaches: (1) using an automated evaluation and (2) manually examined the summaries to see the similarities and diversities of the generated summaries. We hypothesized that a further look at the summarization evaluation would give a more diverse definition of 'what is a good' summary.

1.2.2 The Use of Different Document Types to Support the Summary

In the first research question, we applied our algorithm to news documents, where both the article to be summarized and the supported documents have the same document format.

For the second part of the study, we investigate the use of social media. In this experiment, we applied the same Affinity Graph algorithm using tweets as our source of related document.

1.2.3 Summarization Evaluation: Relevant vs. Preference

For years, automated evaluations (such as ROUGE), have been used to evaluate summary quality. These evaluations focused on the similarity of the content (mostly word similarity) of the reference and automated summaries. The use of crowdsourcing has been successful in producing relevance judgments (Alonso, Rose, & Stewart, 2008). Thus, we assumed that the use of crowdsourcing for user preference in document summarization would provide us a better understanding of what is a good summary.

The result from the summary evaluation experiments (automated and user preference) will be used to demonstrate the most applicable setting for the tweet-biased Affinity Graph approach. The results will also prove if the use of expanded documents is a preferable approach, and thus would be able to improve single document summarization.

1.3 Contribution of the Thesis

By answering the research questions above, we contribute:

1. **Development of an automatic summarization system.**

   In this thesis, an automatic summarization system is developed. The summarization system will be tested using two different datasets, where a new framework will be applied to the new dataset. This is discussed in Section 3.1 for Affinity Graph used in previous work and Section 4.3 for the proposed tweet-biased summarization system.

2. **Investigation of different means and settings to produce summaries.**

   We examined different ways to measure similarity of related documents (e.g. Cosine Similarity, Okapi BM25 and Language Model), different related document settings and a different number of related documents to support the summary. This is discussed in Section 3.4 where the results showed that using different settings and different number of documents gave different ROUGE scores.

3. **Explore the use of tweets as the related document**.

   The Affinity Graph algorithm is applied to a social media dataset. This new framework examined the use of tweets as the related document in the multi-document environment (in Section 4.3). A detailed discussion on the results is presented in Section 4.5.

4. **Development of a summarization system to manually extract sentences and crowdsourcing evaluation system**.

   A manual summarization system to create gold standard summaries for the automated evaluation is also developed and described in Section 4.3.1. A proposed framework of using crowdsourcing to evaluate document summaries is also discussed in Section 5.1.1.

5. **Further analysis of different summary evaluation**.

   The detailed discussion of the summary evaluation is done in Section 3.4, Section 4.3 and Section 5.2. This also includes the qualitative evaluation in Section 5.3 and the discussion on ROUGE-user preference correlation in Section 5.4.

1.4 Guide to the Thesis

The remainder thesis will discuss the automatic summarization system, from the development to the evaluation of the automated summaries.

Chapter 2 presents the automatic document summarization. Chapter 3 will focus on answering the first research question. In Chapter 4, the application of a new dataset to the summarization system is described. Evaluation of the summaries is discussed in Chapter 5. Finally, chapter 6 concludes the thesis with discussions on the contributions and future works.

Chapter 2

Background

In the introduction, we briefly discussed work in document summarization. We also introduced the approach of combining single and multi-document summarization, the area of interest of this thesis. We also briefly discussed the evaluation of summarization.

We will discuss methods and evaluation in document summarization. In Section 2.1, we will discuss the early work and in section 2.2 we will focus on discourse-based graph-based approaches. The affinity graph approach is discussed in Section 2.3, and Social Media Summarization, a current interest in document summarization is discussed in Section 2.4. Section 2.5 will focus on different evaluation methods in document summarization. We will conclude our literature findings in Section 2.6.

2.1 Introduction to Document Summarization

Radev, Hovy, & McKeown (2002) identified four main types of summarization: *Indicative,* help a user decide whether to read a document or not; *Informative* create a shorter, but still detailed version of the document; *Topic- oriented* produce a summary with a particular focus, commonly defined by a user; *Generic* produce a summary with a focus defined by the document's author. However Hahn and

Mani (2000), suggested that a summary can be either an indicative, informative, or critical (adding own opinion in the content).

2.1.1 Single and Multi-Document Summarization

Numerous works in document summarization discussed the use of statistical features that indicate parts of a document that are important or not in generating a summary (Luhn 1958; Edmundson 1969; Rath et al. 1961; Pollock & Zamora 1975). Luhn (1958) introduced a list of words (later called 'stopwords') where he identified their presence as the 'noise' of the document and should be eliminated. Edmundson (1969), Pollock & Zamora (1975) and Rath et al. (1961) used cue phrases and high-frequency words in identifying sentences for their summary.

Later work on machine learning added a new perspective to summarization. The use a Bayesian classifier to identify features in a document (Kupiec, Pedersen, & Chen, 1995) was based on the work done by Edmundson (1969). Lin & Hovy (1997) proposed the use of sentence position to learn individual features of the word and phrases. Myaeng & Jang (1999) used lexical and statistical information from a document corpus where their system was similar to the system proposed by Kupiec et al. (1995). A statistical model was used to select document terms and phrases for a summary (Witbrock & Mittal, 1999). Later the model was used to generate document headlines (Banko, Mittal, & Witbrock, 2000). Conroy & O'leary (2001) used a Hidden Markov Model (HMM) to improve document summarization. They applied similar features used by earlier work, such as the position of sentences and number of terms in the sentences.

In recent years, there has been considerable interest in multi-document summarization. One of the early approaches to such summarization was by McKeown & Radev (1995) who proposed a system for newswire summaries, called SUMMONS. They incorporated templates and extraction rules to better manage domain-specific articles (Radev & McKeown, 1998). Salton, Singhal, Mitra, & Buckley (1997) applied techniques for automatic hypertext link generation to generate a multi-

document summary, where they used a cosine similarity coefficient to link paragraphs within and across documents. Other approaches in multi-document summarization include the use of a graph representation (Mani & Bloedorn, 1999), a vector model in a semantic space of documents (Ando, Boguraev, Byrd, & Neff, 2000) and the use of rhetorical structure theory (Teufel & Moens, 2002).

Hovy & Lin (1999) introduced SUMMARIST, which used topic identification and cue word interpretation to generate a summary. The Centrifuser project used documents from a digital library as its input, and able to identify the similarities and differences of the documents in the produced summary (Kan & Klavans, 2002; K. R. McKeown et al., 2001). Centrifuser. (Lin & Hovy, 2002) introduced NeATS (Next Generation Automated Text Summarization), using techniques drawn from single document summarization (term frequency, topic signature and term clustering). An updated version, iNeATS (Leuski, Lin, & Hovy, 2003) added interactivity allowing users to control parameters (size, redundancy, topic) of the summary.

Harabagiu & Lacatusu (2002) proposed a system called GISTexter, which used three processing stages: sentence extraction, sentence compression, and summary reduction. GISTexter was classified as a Question-Directed Summarization system because of its ability to identify content based on a user's need (Harabagiu, Hickl, & Lacatusu, 2007). MEAD[1] (Radev, Jing, Styś, & Tam, 2004) used a centroid cluster to compute topic characteristics and was also used as a component of NewsInEssence[2] system (Spärck Jones, 2007).

2.1.2 Query-Based and Generic Summarization

Significant work has been conducted in generic summarization. The topics from the documents can be derived by identifying the sentences where we believed that important topics would be comprehensively covered. Methods such as Latent Semantic Analysis (LSA) are used in identifying sentences for generic summarization (Gong & Liu, 2001). Clustering is also another method

[1] http://www.summarization.com/mead/

[2] http://www.newsinessence.com

commonly used in generic summarization. Two different ways of clustering were discussed by Zha, (2002) and Kruengkrai & Jaruskulchai (2003): the former a graph clustering algorithm based on sentence similarity; the later work based on clustering of words in a sentence.

Generic summarization in a multi-document environment is also discussed by Goyal et al. (2013), Hachey (2009), P. Hu, Sun, Wu, Ji, & Teng (2011), Kozorovitzky & Kurland (2009), Kumar, Salim, Abuobieda, & Albaham (2014), Nenkova & Louis (2008) and Wan & Xiao (2010). They applied different methods, from exploiting the inputs and relations of the multi-documents, to machine learning method (i.e. fuzzy reasoning, document fusion and graph-based algorithms). Generic summarization techniques have been applied to both single and multi-document summarization (Alguliyev, Aliguliyev, & Isazade, 2015; Mani & Bloedorn, 1998).

Most of the approaches used in query-based summarization applied machine learning methods or knowledge-based system extraction, in order to focus a summary towards a users' query. The use of a graph to define the relationship between a query and a document was proposed by Bhaskar & Bandyopadhyay (2010), Bosma (2005), Jagadeesh, Pingali, & Varma (2007) and Varadarajan & Hristidis (2005, 2006). The use of an ontology was in the medical domain was proposed by Mollá (2010) and Ping & Verma (2006). Use of a Bayesian Statistical Model was proposed by Daumé & Marcu (2006).

2.1.3 Extraction and Abstract Summarization

There are two approaches to automatic text summarization: extraction and abstraction (Hahn and Mani, 2000, Spärck Jones, 1993, Nenkova and McKeown, 2011). Extraction methods form summaries from text extracted from the document(s) to be summarized and have been found to be easier create. Abstraction adds in the process of paraphrasing or writing from scratch sections of the document, which is considered a more difficult approach. Abstraction requires a more semantic

understanding of the source text. As emphasized by Silber & McCoy (2000), such understanding requires some form of specific (or domain) knowledge base.

In this thesis, we will focus on extraction based techniques.

2.2 Document Summarization Approaches

In single and multi-document summarization, there are different approaches used to extract, rank, and select the sentences that are considered most relevant to a summary. We discuss the discourse-based and graph-based approaches, which are both common.

2.2.1 Discourse-Based Approaches

The discourse-based approach typically involves three stages (Mani, 2001). Initially, an exploration of document structure takes place before assessing sentences in the next stage. Finally, the summary is generated by extracting relevant sentences. Rhetorical Structure Theory (RST) is commonly used to identify document structure (Mann & Thompson, 1988), see also (Bosma, 2005; Carlson, Marcu, & Okurowski, 2003).

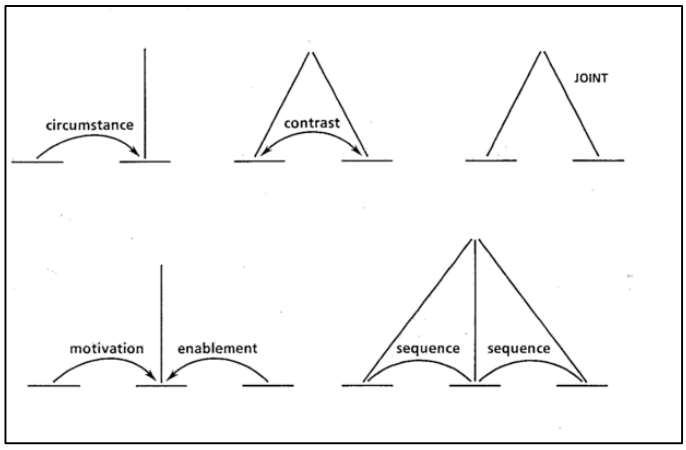


Figure 2.1: The five schema types (Mann & Thompson, 1988)

In RST, a structure is created to represent relations between sentences. The structure was based on the five schema types as introduced by Mann & Thompson (1988), see Figure 2.1. The schema represents a structural analysis of sentences.

The RST approach has been widely used, especially in single document summarization. Structural analysis formed the basis of sentence weights in (Bosma, 2005).While in (Marcu, 1997a, 1997b), they applied RST to identify important units in a document. Teufel & Moens (2002) proposed an RST-based summarization system for scientific articles, identifying seven rhetorical categories. In (Bosma, 2005), they applied RST to create a graph representation of a document from which query-based summarization was produced. They also found that their method can be applied to a non-RST graph-based approach. This is because their method used two graphs: an RST one to link sentences, and another to extract sentences.

The semantic and rhetorical relationships of sentences within a document were captured and combined (Atkinson & Munoz, 2013). RST was applied and a corpus-based analysis was used in a web-based multi-document summarization framework. Combining the two approaches resulted in summaries that were more accurate than the state-of-the-art.

Because of the complexity of RST (due to its need to analyse the complex semantic representation of the document), we believed that it would not be a suitable approach for our proposed framework. This is because most RST systems rely on ontologies or language corpora, resulting in a summarization system that is slow and with limited coverage of many domains. We believed that a non-semantic approach would be much cheaper and less complex than RST.

2.2.2 Graph-Based Approaches

In a graph-based summarization system, document sentences are represented as nodes (represented as numbers in the graph, see Figure 2.2) connected by edges that are weighted to represent inter-sentence

similarity (Thakkar, Dharaskar, & Chandak, 2010). The more similar two sentences are, the higher the edge weight. In our figure, the closer the nodes are to each other, the stronger the sentence similarity.

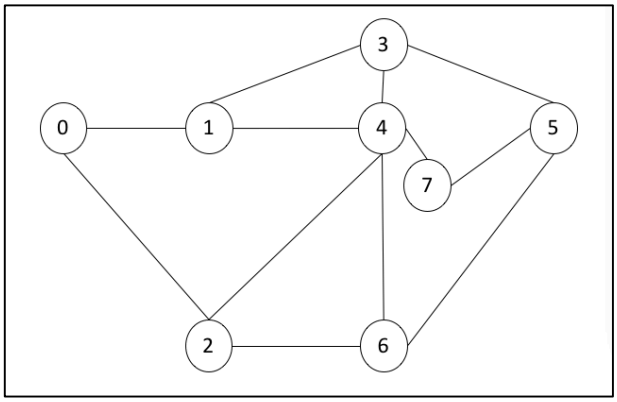


Figure 2.2: A Graph

Mani & Bloedorn (1997) used a graph-based algorithm for multi-document summarization. They built a graph representation to identify relationships between documents to generate a summary. They later improved their graph-based approach (Mani & Bloedorn, 1999), identifying relationships between sentences within a document and its related documents. Yoo, Hu, & Song (2006) and Plaza, Díaz, & Gervás (2011) proposed a semantic graph-based summarization approach in a medical domain, and Li, Du, & Shen (2013) used a graph-based algorithm for sentence ranking and applied it to an update summarization system.

Giannakopoulos, Karkaletsis, Vouros, & Stamatopoulos (2008) proposed the use of N-gram graphs to evaluate summaries. Even though they showed that their evaluation method was comparable with other automated summarization methods, the complexity of the approach limited its take up by others.

We believed that a graph based approach would be able to capture local and global relationships of both documents and sentences. Another advantage of graph is that the nodes-edges can be represented by similarity or semantic relations (Sizov, 2010). Thus, we can use many relations to connect the documents, and it is easy to measure a range of similarity scores (pairwise similarity) between document.

2.3 Affinity Graph Approach

The use of Affinity Graph was based on work by Zhang et al. (2005), where they improved a document ranking by investigating the *diversity* and the *information richness* of documents. The authors revealed that their Affinity Graph approach significantly improved ranking as tested on datasets from Yahoo!, Open Project Directory, and a from newsgroups. They compared their method with a K-Means clustering algorithm reporting significant improvement over the baseline.

Consequently, researchers believed that an Affinity Graph approach would be able to improve document summarization by including information from related documents. The use of an Affinity Graph for summarization was first discussed by Wan & Xiao (2010) who used the algorithm for summarization and keyphrase extraction (Figure 2.3).

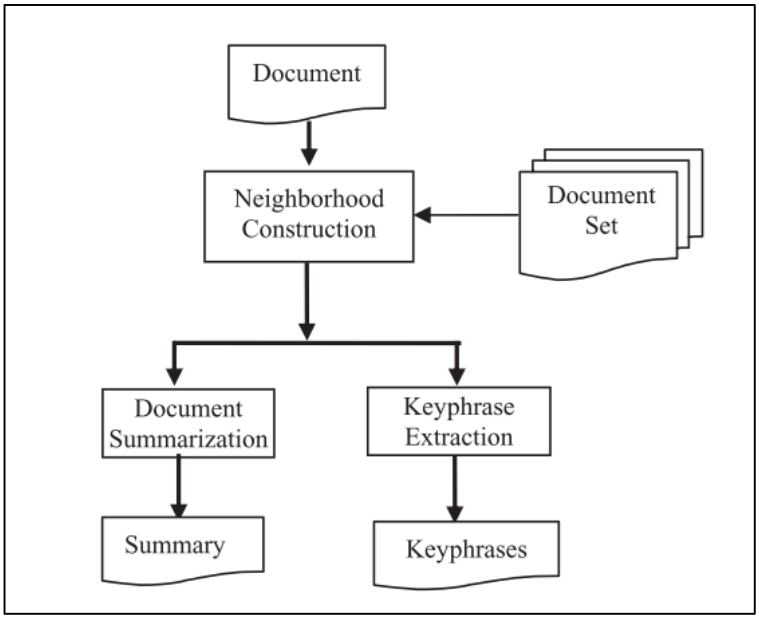


Figure 2.3: Wan & Xiao (2010) proposed framework

They constructed a neighbourhood of related documents using a cosine similarity measure. In document summarization, they defined the relationship of the document to be summarized and the related documents using a *confidence value*; the more similar the documents were to each other, the higher the confidence. Here, Wan & Xiao (2010) applied three steps:

1. Neighborhood-level sentence graph building,
2. Neighborhood-level sentence evaluation, and
3. Document-level redundancy removal

Step 1 defined the Affinity Graph algorithm. In Step 2, Wan & Xiao (2010) ranked sentences based on an informativeness score (*if_score*) derived from the algorithm. In Step 3, they used a greedy algorithm from Zhang et al. (2005) - a variant of Maximal Marginal Relevance (MMR) from Carbonell & Goldstein (1998) - to penalize the scores of the sentences that highly overlap with other informative sentences.

Their approach showed the complexity of the computation in finding the related documents, where the document to be summarized was compared with each document in the collection: a high computational cost. The cost was made greater by the number of sentences in the documents to be ranked: $(k+1)\times Ave_{sentence}$ (where $(k+1)$ is the number of related documents and $(Ave_{sentence})$[3] is the average number of sentences in the expanded document collection). The same computational complexities were also applied to the sentences of the related documents. Thus, Affinity Graph showed a complex relationship as more documents were added to the graph.

More discussion on the work by Wan & Xiao (2010) is in Chapter 3.

[3]For single document summarization, the number of sentences to be analyzed is $(Ave_{sentence})$.

Context-Based Indexing for Document Summarization (2013)

Goyal et al. (2013) improved summarization by proposing a context-sensitive document indexing using a Bernoulli model of randomness to develop a graph-based sentence ranking algorithm. They also presented three hypotheses of their approach:

(1) A summary evolved around a *topical term*,

(2) The topical term appears more than the non-topical terms across the document to be summarized and all related documents

(3) A graph is created using lexical association to improve summarization accuracy.

Goyal et al. (2013) suggested a more complex use of the Affinity Graph, where they explore the use of lexical association in their proposed method. In their paper, they calculated the probability of a term (*t*) appearing in a document. They applied a Bernoulli model to calculate the distribution of *t* and use it as the input for their sentence similarity. Hence, their work combined lexical and graphical summarization methods.

Goyal et al. (2013) also used the same definition of terms as Wan & Xiao (2010). Both researchers conducted their experiments using the same dataset and evaluation method. Both work will be described and discussed more in the next chapter as their work will be the basis of our proposed framework.

Collaborative Approach Using Social Contextual Information in Sentence Ranking (2011)

We also found similarities with the work by P. Hu, Ji, Sun, Teng, & Zhang (2011a) and P. Hu, Sun, et al. (2011), who focus on building a social context summarization using user's tag in a social bookmarking website[4]. Here tags and tweets from user's were exploited to identify more relevant information and thus improve document summaries. P. Hu, Ji, et al. (2011a) applied the Affinity

[4] https://delicious.com/

Graph algorithm to rank documents' sentences by considering the user context from the tagging. They reported that the proposed approach (called *SocialContextSum*) significantly improved over baselines.

With the increase of news articles and social media websites, different types of documents (tweet, tags, forums, blogs) have emerged that might be exploited for document summarization. Of particular interest are tweets - that we believe might provide more meaningful information than tags.

2.4 Social Media Summarization

Nenkova & McKeown (2011) anticipated that social media will be a new area of interest in document summarization, due to the growth of popularity in social networking. Blogs, comment, tweets and social bookmarking (tagging) are different types of social media documents.

In blog comment summarization, M. Hu, Sun, & Lim (2007, 2008) identified important sentences to be extracted (M. Hu et al., 2007), and later used comments to understand user feedback. Parapar, López-Castro, & Barreiro (2010) also used blog posts and comments to generate blog snippets. While these work exploited the blogs comments to generate better summaries, S Mithun (2010), Shamima Mithun & Kosseim (2009) and Xiaodan Song, Chi, Hino, & Tseng (2007) used the blog posts entries to create summaries in their blog summarization system.

The use of tagging and tweets in summarization involved is more challenging due to the limited information in the content. Boydell & Smyth (2007), P. Hu, Sun, et al. (2011), Park, Fukuhara, Ohmukai, Takeda, & Lee (2008) applied their summarization techniques to social bookmarking websites. These works used the tags as content clues to score and rank sentences in web pages. Each work found their proposed summarization system benefited from the 'tags'.

There are two summarization types discussed in tweets summarization. Kothari, Magdy, Darwish, Mourad, & Taei (2013), Mackie, McCreadie, Macdonald, & Ounis (2014), Nichols, Mahmud, & Drews (2012), Ritter, Cherry, & Dolan (2010) and Sharifi, Hutton, & Kalita (2010)

applied their methods to summarize tweets themselves, where Gao, Li, & Darwish (2012) used tweets to jointly summarize a web document. P. Hu, Ji, et al. (2011b), P. Hu, Sun, et al. (2011) and Yulianti, Huspi, & Sanderson (2015) described a summarization system that incorporated tweets in a single document summarization. Štajner et al. (2013) evaluated different methods to conduct news selection from tweets, where they identified interesting messages from social media related to news articles.

Mackie et al. (2014) aimed to compare different evaluation measures for microblog summarization, applying three different systems (Centroid, SumBasic and Hybrid) to summarize tweets from four microblog datasets. P. Hu, Ji, et al. (2011b) applied an Affinity Graph summarization approach to rank document sentences based on the social context identified from the tags.

Tweet-Biased Summarization (2015)

In the work of Yulianti (2013) and Yulianti et al. (2015) they proposed a tweet-based summarization system and developed a new evaluation dataset. They were inspired by Parapar et al. (2010) who used blog comments to generate snippets for the blog search results. Their results showed an improvement up to 32% compared to a baseline.

Yulianti (2013) extracted tweets from a microblog dataset and selected those that had links to a set of web documents. They found 493 documents with a minimum of 10 linked tweets. The main contribution from this work is the development of a Tweet-Biased Summarization (TBS) system and a Generic Summarization (GS_{sn}) that used only information from the local document.

TBS is based on the ranking of the tweets that link to a document. In the first part of generating a TBS, the tweets were ranked based on their relevance to the document. Here, they selected the top 30% of the ranked tweets and defined them to be the ‘novel tweets’. These tweets were then combined to form a new query, to be used in the second part of the process.

In the second part, the process was repeated for sentences from the web document. This time, the sentences were ranked based on the new query (from the first part), and then reapplying a novelty detector system. The system was applied to filter redundant content. The novelty score was the cosine similarity between the unique terms in the document and the retrieved tweets/sentences of the document. In the second part, the novelty detector system was used to re-rank the sentences and generate a summary.

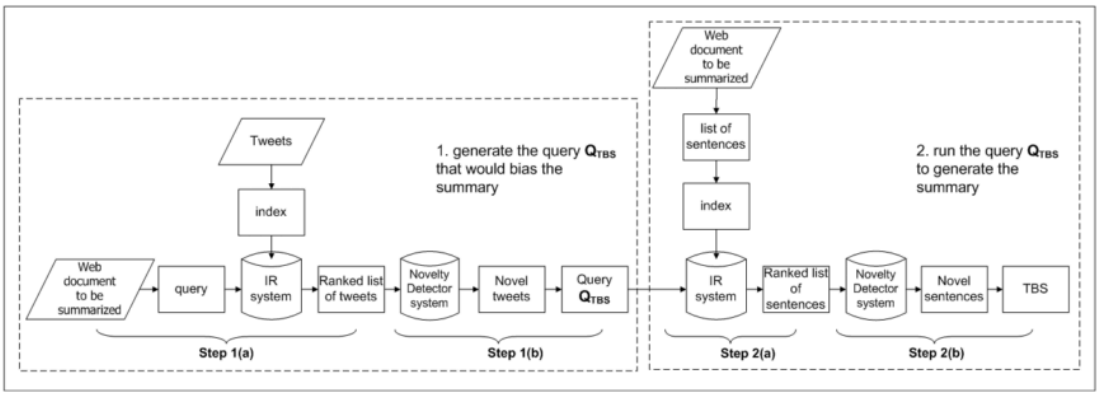


Figure 2.4: The TBS system (Yulianti, 2013; Yulianti et al., 2015)

Figure 2.4 showed the framework, where we noticed that the sentences from the single document go through the novelty detector algorithm twice. First to rank the sentences based on the query (to identify the 'related' sentences based on the tweets), and secondly, to rank the sentences based on the *novelty score* to create a summary. Even though the two-time ranking process is questionable, Yulianti et al. (2015) explained that the process was done to replicate Parapar et al. (2010)'s work and they believed it would not disturb the whole summarization process.

We believed that our proposed Affinity Graph algorithm would provide a better solution in the same environment discussed by Yulianti et al. (2015) for document summarization. The detailed comparison results by Yulianti et al. (2015) is discussed in Chapter 4 and Chapter 5.

2.5 Summarization Evaluation

Evaluating a summary is subjective, and the criteria of a 'good' summary depend on the purpose a summary should serve.

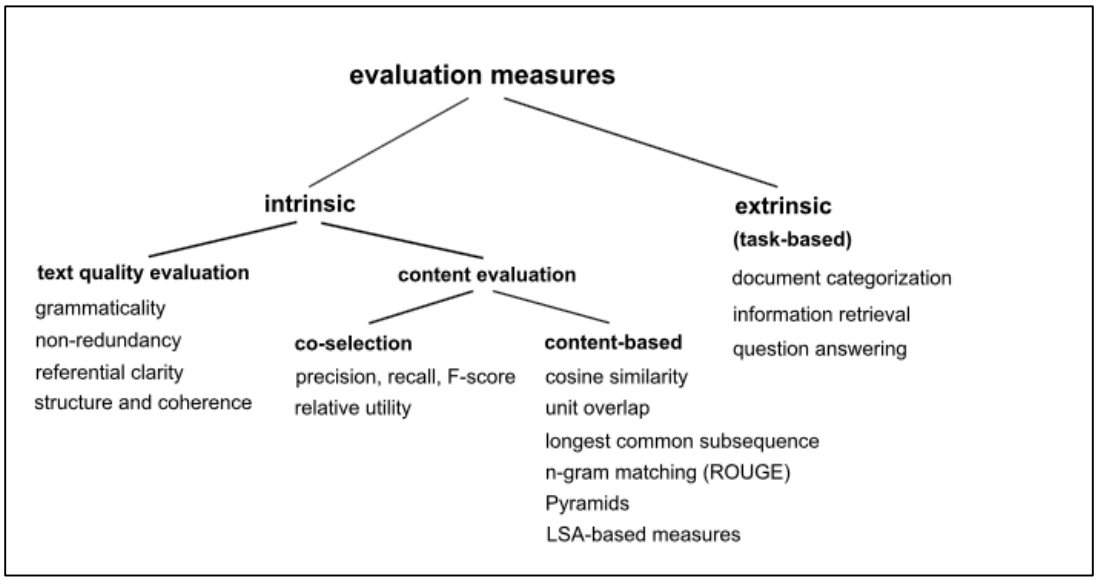


Figure 2.5: Summary evaluation measures (Steinberger & Ježek, 2012)

As shown in Figure 2.5, the two main summary evaluation measures are intrinsic and extrinsic (Steinberger & Ježek, 2012). Intrinsic evaluation measures the summary based on its reliability compared to its source document or a summary produced previously. Extrinsic evaluation measures the outcome of a summary based on specific tasks, and it varies depending on different systems. Relevance judgment is one example of an extrinsic measure, and this evaluation is used to judge whether a document/summary is accurate. In this chapter, we will focus on past evaluation work discussed in the thesis.

2.5.1 Intrinsic Evaluation

Intrinsic evaluation aims to evaluate the quality and informativeness of a summary (Mani, 2001) by comparing the generated summary to a human generated 'model' summary. We have an interest to

use content evaluation to evaluate our summarization system. This is because it is important for us to know if the auto-generated summary that we have created are similar to an 'ideal' summary. The common evaluation is by calculating the *recall* and *precision* of the summary, see Figure 2.6.

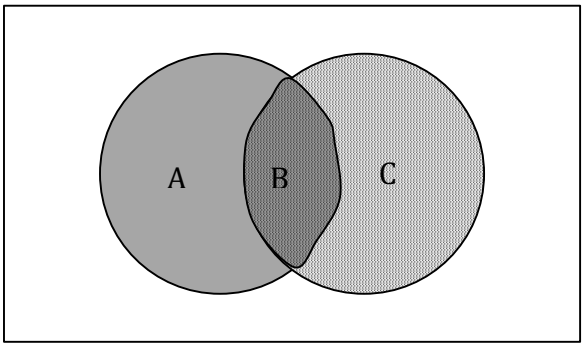


Figure 2.6: Example of Recall and Precision

Let say A is the gold standard summary, C is the system-generated summary and B is the overlap of sentences in A and C. Recall calculates the ratio of sentences in the generated system (B) that is in the gold standard summary (A):

$$Recall = \frac{B}{(A+B)} \quad (2.1)$$

As for precision, it is defined by:

$$Precision = \frac{B}{(B+C)} \quad (2.2)$$

where precision calculates the sentences that overlap with the gold standard and the system-generated summary (B) as compared with the whole system-generated summary (C).

Nenkova & McKeown (2011) discussed the weakness of the measures. First, there is a likelihood of *human variation* in generating the gold standard summary; because of the possibility that a system extracted good sentences, but due to small overlap with the gold standard summary, the recall/precision score is much less than it perhaps should be. In the second problem, two summarizers might each extract a different sentence that both appeared in the gold standard summary. The recall/precision score would be the same but the summary would be very different.

In describing other evaluation approaches, we introduce the **Bi**l**L**ingual **E**valuation **U**nderstudy (BLEU). BLEU was proposed by Papineni, Roukos, Ward, & Zhu (2001), where it measures the *precision* of a machine translated text. The aim of BLEU is to measure the 'closeness' between a human translation and a system-generated translation. BLEU was defined as a modified n-gram precision measure (Papineni et al., 2001) and computed as:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \quad (2.3)$$

where:

$Count_{clip}(n-gram)$ is the maximum number of n-grams co-occurring in a candidate document/sentences and a reference document/sentence; and
$Count\ (n-gram)$ is the number of n-grams in the candidate sentence.

Papineni et al. (2001) also introduced the *sentence brevity penalty* (BP), where the penalty score is used to make sure a high score candidate has equal length and word selection compared to the reference document/sentences:

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \quad (2.4)$$

where:
c is the length of the candidate document/sentences; and
r is the length of the reference document/sentences.

Thus, BLEU is defined by:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (2.5)$$

where:
N is the length of n-grams; and
w_n is a weighting factor.

We believed that the BP gives the main advantage of BLEU; all systems are treated equally despite having different styles of human reference. However, Xingyi Song, Cohn, & Specia (2013) claim that BLEU does not work well at the sentence level, a problem addressed by many (Callison-Burch, Osborne, & Koehn, 2006; Xingyi Song et al., 2013).

Inspired by the work in machine translation evaluation, a content-based evaluation approach was proposed: ROUGE, or **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation (Lin, 2004). In the ROUGE toolkit, five evaluation metrics are available:

- ROUGE-n: n-gram based co-occurrence statistics.
- ROUGE-L: Longest Common Subsequence (LCS), which calculates sentence level structure and identifies the longest co-occurring sequence of n-grams.
- ROUGE-W: this is the Weighted LCS-based statistics that support consecutive LCS.
- ROUGE-S: skip-bigram (any pair of words in their sentence order) based co-occurrence statistics
- ROUGE-SU: skip-bigram and unigram-based co-occurrence statistics.

We can see that ROUGE-L and ROUGE-W measure the LCS that is shared by the candidate and reference summaries, however ROUGE-W gives weights to consecutive matches in the candidate sequence. An example on this was discussed by Sizov (2010) as follows:

Reference Summary	:	***the white cat*** went missing
Candidate 1	:	this is because ***the white cat*** was hungry
Candidate 2	:	***the*** man in ***white*** kicked a ***cat***

In this example, Candidate 1 and 2 would get the same ROUGE-L score, as [*the*, *white*, *cat*] appears in both Candidate 1 and 2. But Candidate 1 will get a better score for ROUGE-W, as [*the*, *white*, *cat*] did not appear in as a consecutive match in Candidate 2.

For ROUGE-S, the Candidate 1 has only *one* reference biagram [the gunman], but *six* skip-bigram [police, killed], [police, the], [police, gunman], [killed, the], [killed, gunman] and [the, gunman]:

Reference Summary	:	police killed ***the gunman***
Candidate 1	:	***the gunman*** killed police
Candidate 2	:	***gunman*** the killed police

Thus, ROUGE-S measures the overlap ratio between a candidate summary and a set of reference summaries (Lin, 2004). But in ROUGE-SU, it included the unigram[5] overlap in the candidate summary. This is to solve the issue that if in a candidate summary, it does not have any sentences with any word pair in the reference summary (as in Candidate 2 – it would get zero score in ROUGE-S).

The most used ROUGE evaluation metric is ROUGE-n, where it calculates the overlap of n-grams between candidate and reference summaries; note it is possible to have more than one reference summary. (More detail on ROUGE-n is found in Chapter 3).

Lin (2004) found that ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S are best used for single document summarization. While ROUGE-1, ROUGE-L, ROUGE-W, ROUGE- SU4, and ROUGE-SU9 are best when evaluating short summaries.

Another method for intrinsic summary evaluation is The Pyramid Method proposed by Nenkova & Passonneau (2004) who claim it is a more reliable evaluation due to the ability to get sentences using different words but with the same meaning (which they called Summary Content Unit – SCU).

[5] represented by the U in ROUGE-SU.

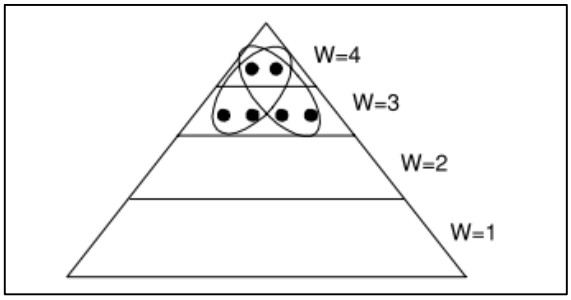


Figure 2.7: A Pyramid with 4 SCUs (Nenkova & Passonneau, 2004; Nenkova, Passonneau, & McKeown, 2007)

The idea is to get as many SCUs in manual summaries so that the units will get higher weights and be placed on top of the pyramid (Figure 2.7). A pyramid represents the number of SCUs for the summaries, where each sentence was indexed by the position of its appearance in its respective summary. In the example discussed by Nenkova & Passonneau (2004):

A1	In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.
B1	Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.
C1	Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.
D2	Two Libyan suspects were indicted in 1991.

The first column (the alphabet) represents the respective summary (in this case there are four different summaries), and the number represents the position of the sentences in its summary. From this example, they have obtained two SCUs:

SCU1: *two Libyans were officially accused of the Lockerbie bombing*
From A1: [two Libyans] [indicted]
From B1: [Two Libyans were indicted]
From C1: [Two Libyans,] [accused]
From D2: [Two Libyan suspects were indicted]

SCU2: *the indictment of the two Lockerbie suspects was in 1991*
From A1: [in 1991]
From B1: [in 1991]
From D2: [in 1991.]

This will give SCU1, the weight 4 and SCU2 (weight 3), thus SCU1 will be placed on top of the pyramid, and SCU2 in tier 2. At the end of the annotation procedure, the pyramid will have tiers that contain SCU with the same weight. The evaluation of each summary is done calculating the ratio of the weighted sum of its SCUs to the sum of the peer (reference) summary (with the same number of SCUs). Thus, the score represents how much information (SCUs) appears in both summaries.

Even though the method is semantically driven, the task is costly and involved a large use of human labour (Nenkova & McKeown, 2011). As the Pyramid Method is developed for abstractive summaries (Nenkova & McKeown, 2011; Nenkova et al., 2007), we believed this evaluation might not be suitable for our approach, which is more to an extractive summary approach.

2.5.2 Extrinsic Evaluation

Extrinsic evaluation judges the quality of a summary by assessing how well it can assist humans to complete a specific task (relevance decision). One example of an extrinsic evaluation is the document relevance judgment: can a human judge relevance reading a summary just as accurately as if they had read a full document.

There have been several extrinsic evaluations in document summarization. In Figure 2.5, Steinberger & Ježek (2012) included three types of extrinsic evaluation: *Document Categorization*, *Information Retrieval* and *Question-Answering (Q&A)*. Mani (2001) categorized extrinsic evaluation into four categories: Relevance Assessment, Reading Comprehension, Presentation Strategy Evaluation and Mature System Evaluation. In the last two categories, both evaluations were reported hard to be applied as both evaluation involves human factor studies. Here, subjective features, such as

presentation (colours, iconology etc.), quality of the summary solutions and user satisfaction is measured.

The quality of a summary is usually presented in how good can the summary be used to categorized a document, even without reading the whole documents. Mani et al. (1999) used extrinsic evaluation in the TIPSTER Text Summarization Evaluation (SUMMAC) in two task: ad hoc and categorization task. In the categorization task, the generic summary was evaluated on whether it has enough information to allow the participants to categorized a document. Argumentative Zoning, a rhetorical classification task proposed by Teufel (2000), was used by Siddharthan & Teufel (2007) in their work to categorized scientific documents. In the later work, they found that their method showed a higher human agreement (Kappa) in the extrinsic evaluation compared to the work from Mani et al. (1999).

Another extrinsic evaluation type discussed by Mani (2001) was the Reading Comprehension, where the task requires for a human to fully read and understand a summary. Here the human reading comprehension is tested, where a set of questions were asked to see if he/she would be able to accurately answer them. If the percentage of the correct answer is high, it is assumed that the summary is highly informative. The Question-Answering task is based on this task. One main reference work for this task was performed by Morris, Kasper, & Adams (1992). Teufel (2001) also performed the question-answering task, where they identified that keyword, random sentences, and abstracts do not provide enough information to complete the task.

We noticed that the Information Retrieval task has a similar characteristic with the Relevance Assessment (Mani, 2001). One definition of relevance is the measure of correspondence existing between a document and a query as determined by the users (Saracevic, 2007). Most summaries are generated based on the assumptions that it is topically relevant to the document. Whereas, the human judgements are based on the *internal* and *external* context of the users. *Internal context* concerns on the user's knowledge, feelings, and expectations; and the *external context* considers the user's task

and their environment. A user's judgement is dynamic and it is based on many document attributes (topic, clarity etc.); hence the relevance judgement varies across users.

Most work applied relevance judgement task, where they asked human evaluators to judge if the summary is *relevant* or *not relevant* to a given query/topic (Bonnie Dorr, Christof Monz, Douglas Oard, Stacy President, David Zajic, 2004; Brandow, Mitze, & Rau, 1995; Jing, Barzilay, McKeown, & Elhadad, 1998; Mani et al., 1999; Mani & Bloedorn, 1997). Current relevance judgement task is performed online, where a group of man power are used to evaluate document summaries (Mackie et al., 2014; Yulianti et al., 2015). This evaluation method has become popular due to a large number of results can be obtained in a short time. A further discussion on this evaluation method is discussed in Chapter 5.

2.6 Conclusion

We found that there are many work has been done in all aspect of document summarization. With the different types of document summarization, our interest is to explore the possibility of using related documents to improve generic-extraction single document summaries. Discourse approach has the advantage to semantically extract the content of a document; this would be a better solution for abstractive summaries. A graph-based approach is more applicable to our proposed summarization method because the graph-based approach can identify the relationships of sentences between documents without having to initially create the relationships between the documents.

The necessity to evaluate how a good a summary leads to the discussion on different summarization evaluation. However, the choice of using intrinsic or extrinsic evaluation depends on the goal of the summarization systems. While intrinsic evaluation is much recommended, extrinsic evaluation that involved 'real' users has also become more important.

Chapter 3

Re-examining Affinity Graph for Document Summarization

The previous chapter had discussed different types of document summarization approaches and their evaluation. Kozorovitzky & Kurland (2009), Wan & Xiao (2010) and Goyal et al. (2013) discussed the potential of generating a summary that combines single and multi-document summarization. Thus, the concern of this work is to explore the possibilities of this framework and apply it to different document types.

Thus, the first research question (RQ1) is:

"How effective are graph-based algorithm approaches in improving single document summarization?"

We are interested to discover ways to improve single document summarization, and it is believed that the use of 'neighbourhood' documents could improve summarisation by providing more information to sentence selection processes. From the literature review, we discovered that there is an interest in using an Affinity Graph for document summarization: a graph-based ranking approach that examines the relationship between the sentences of related documents. In Goyal et al. (2013) and Wan & Xiao (2010), they showed that the Affinity Graph approach was able to significantly improve a baseline single document summarization system. However, in both works, there was little discussion

on how best to configure the Affinity Graph algorithm. We believed that further discussion and comparison of the generated summaries should be examined in order to understand how best to use information from related documents. This chapter describes this work. It is split into two sets of experiments.

The first set will explore the use of different similarity measures: Okapi BM25, Cosine and Indri Language Model. The measures will be used to identify the related documents and to give similarity scores to each sentence from the document. We also investigate different types of document setup, and explore different possible summarisation parameters. The second set of experiments explore the optimal settings of related documents to be used in the Affinity Graph approach. Hence, the experiments will test the approach by using different number of related documents and also different versions of document length.

Thus, the aim for this chapter is to examine the Affinity Graph approach and to re-evaluate the summaries generated using the Affinity Graph algorithm. As for the evaluation, we will not only discuss on the automated evaluation (by comparing with a different baseline and previous work), but also to manually examined the content of the summaries.

We defined the following terms, which will be used throughout this chapter:

- Local Document: this is the document that we generate our summary for (or the document query).
- Expanded Document: the documents that are related to the Local Document.

3.1 Background Work

As discussed in the previous chapter, graph-based approaches are able to represent the relationships of sentences between documents (Lloret & Palomar, 2012; Steinberger & Ježek, 2012). Erkan & Radev (2004) introduced the concept of the centrality of a sentence to a document. A graph-based

summarization approach was able to produce better summaries compared to a word or sentence level summarization as discussed by Wolf & Gibson (2004).

Graph-based algorithms have been used successfully for web search. Here, documents were given ranks based on their similarity value, locally and globally within the graph. PageRank (Page, Brin, Motwani, & Winograd, 1999), LexRank (Erkan & Radev, 2004) and TextRank (Mihalcea & Tarau, 2004) are examples of graph-based ranking for search and text processing. Mihalcea (2004) also used TextRank for text summarization and demonstrated it using a Document Understanding Conference 2002 (DUC02) dataset. Zhang et al. (2005) introduced an Affinity Graph (AG) algorithm to rank web documents, by optimizing two metrics: diversity and information richness. Wan & Yang (2006) then explored the use of an Affinity Graph (AG) for multi-document summarization and later proposed CollabSum (Wan, Yang, & Xiao, 2007).

3.1.1 Affinity Graph

The use of the Affinity Graph for document summarization was first discussed by Wan & Xiao (2010). They constructed an Affinity Graph (AG): a neighbourhood of documents related using a cosine similarity measure.

We applied the Affinity Graph approach first by using similarity search techniques to identify expanded documents and, next, to calculate affinity values between each of the paired sentences. With the graph in place, we calculate '*informativeness*' scores (which are called *if_score*) to identify important sentences from the local document, which we then extract to form a summary.

To identify the expanded documents, we applied pairwise similarity measures to calculate the pairwise relationship between the local document and the expanded documents (Figure 3.1).

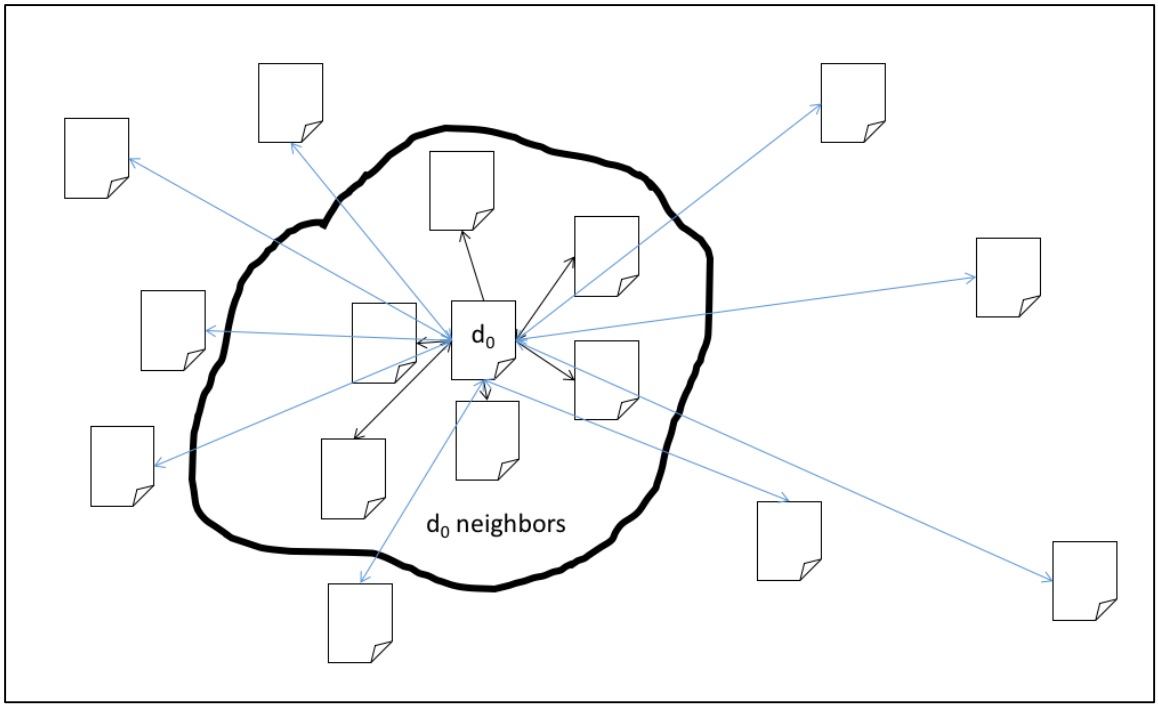


Figure 3.1: Document d_0 (local document) with its neighbours.

The Affinity Graph in Figure 3.1 showed the relationship between document d_0 and its neighbours, represented as a link with an associated 'affinity value'. The documents with black arrows have a higher 'affinity value' and form the neighbourhood of expanded documents (the black line region).

We assumed that the expanded documents described topics that are similar to the local document. Thus, the affinity graph maps relationships between documents and gives scores to show the strength of the relationships between the documents.

Next, the local and expanded documents are split into sentences and the similarity between each sentence of the local document and the expanded documents is calculated (in Part 1b in Figure 3.2).

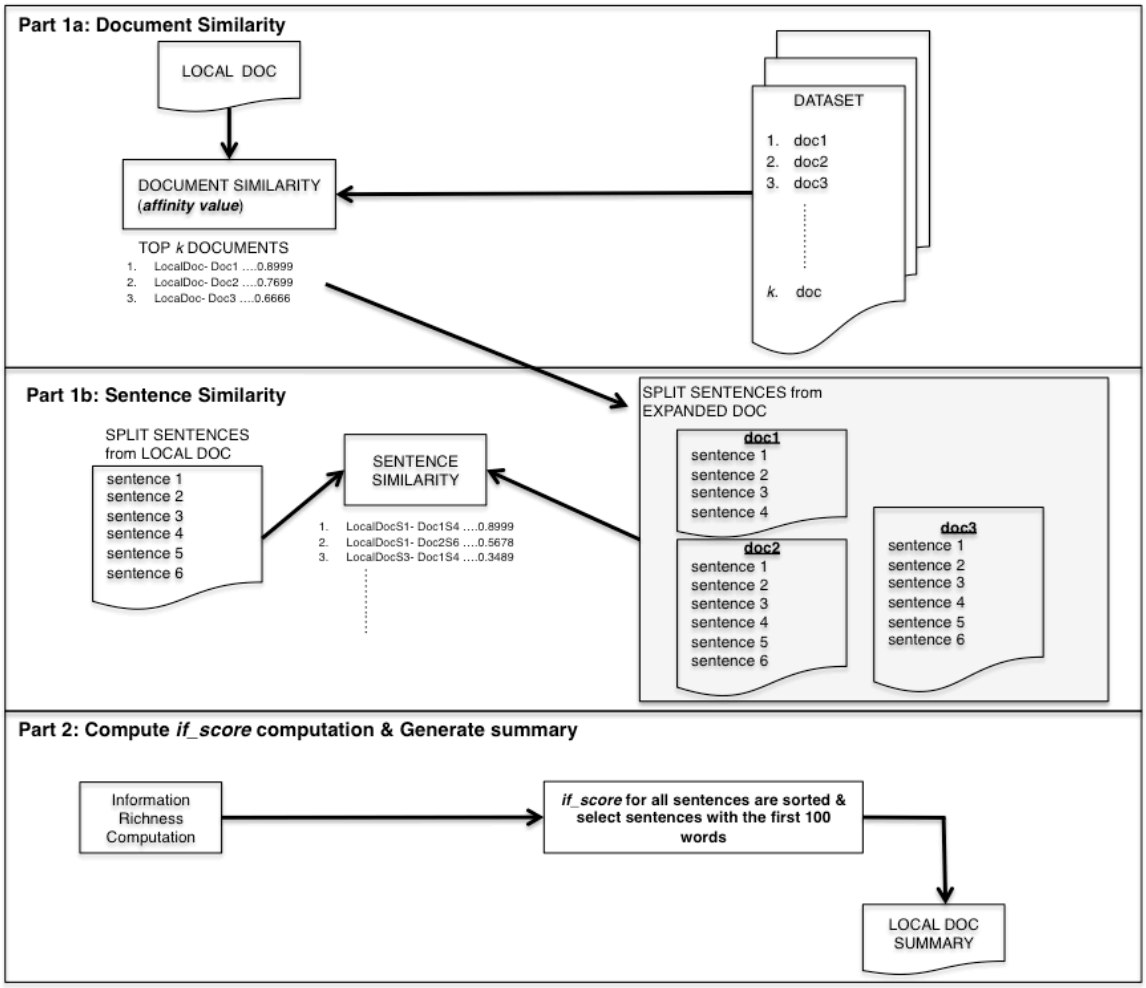


Figure 3.2: The Affinity Graph approach

For our experiments, we evaluated three similarity techniques: Okapi BM25, Cosine and Indri Language Modelling. We used the Lemur project toolkit[1] to calculate similarity values. The features provided by the Project are:

- INDRI Search Engine: used to calculate the similarity of pairs of documents and sentences using Okapi BM25 and Indri Query Language.
- LEMUR toolkit: used to calculate the similarity of pairs of documents and sentences using Cosine Similarity.

[1] http://www.lemurproject.org/

The similarity values of pairwise document and sentences will then be stored in a matrix, M. The sentences from the local document are defined by (s_i) and sentences of the expanded document are defined by (s_j):

$$M_{i,j} = \begin{cases} \lambda \times sim_{sen}(s_i, s_j), & i \neq j, \\ 0 & otherwise \end{cases} \quad (3.1)$$

The matrix $M_{i,j}$ will give the result of a set of scores that represent the importance of each sentence in the affinity graph. Here, we define that if the sentences were within the Local Document (within document link), λ is set to 1. Otherwise, the λ is set to the affinity value calculated from document pairwise similarity calculated earlier. The function $sim_{sen}(s_i, s_j)$ is the similarity between sentence s_i and s_j .

Next M is normalized to $\widetilde{M}$ (see Equation 3.2) to ensure that total of each row [$s_i s_j$] comes to one. This is done because we are interested in the internal structure of the relationship between the sentences. Thus the same range of the relations show how 'related' the sentence based on higher score of the relationship.

In the following equation, S is the set of sentences in the local and/or expanded document set, and $\sum_{j=1}^{|S|} M_{i,j}$ is the total value of the matrix M for the sentence set the document settings:

$$\widetilde{M}_{i,j} = \begin{cases} M_{i,j} \Big/ \sum_{j=1}^{|S|} M_{i,j}, & \sum_{j=1}^{|S|} M_{i,j} \neq 0 \\ 0, & otherwise, \end{cases} \quad (3.2)$$

In Part 2 (Figure 3.2), the scores from the normalized matrix $(\widetilde{M})$ will then be used to calculate the informativeness score (*if_score*) of each sentence for the local document (d_0) by applying Equation 3.3. The *if_score* represents the importance of sentences in documents, the higher the score, the more important.

$$if\_score_{all}(s_j) = \mu \cdot \sum_{all j \neq i} if\_score_{all}(s_j) \cdot \widetilde{M}_{j,i} + \frac{(1-\mu)}{|S|} \quad (3.3)$$

Here, we define:

- μ as a damping factor,
- $\sum_{all j \neq i} if\_score_{all}(s_j)$ is the sum of if_score values for sentence s_j,
- $\widetilde{M}_{j,i}$ is the normalized matrix as in Equation (3.2), and
- $\frac{(1-\mu)}{|S|}$ is the probability that the information flows into any document in the collection.

The $\frac{(1-\mu)}{|S|}$ component is similar to the Markov Chain theory, where information flows between document nodes at each iteration (Zhang et al., 2005). The red line in Figure 3.3 represents the flow of the information of the document, where it can be used in one of the documents ($[\mu]$) or used to any random documents in the collection ($[(1-\mu)]$, which is represented by the green line.

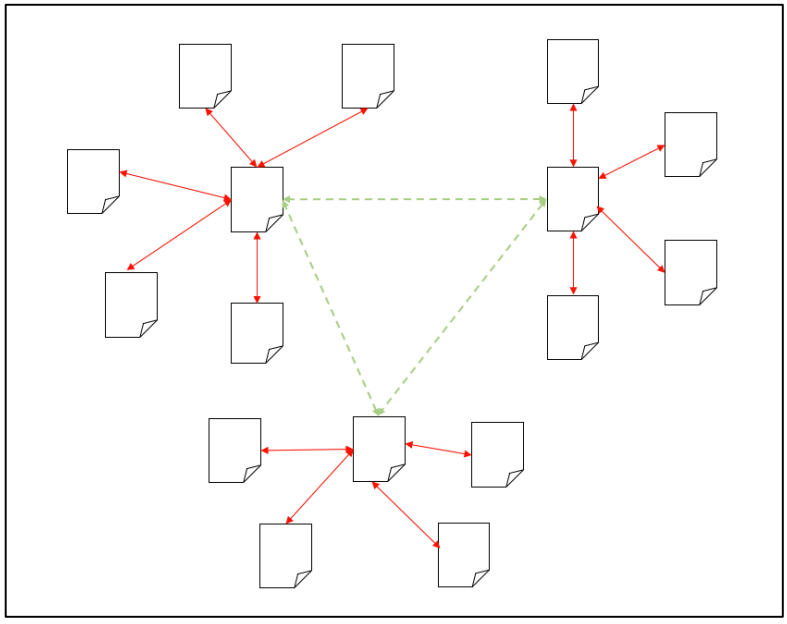


Figure 3.3: Information flow in Affinity Graph (Zhang et al., 2005)

In this experiment, we set all threshold values in line with values chosen in (Wan & Xiao, 2010), with $\mu = 0.85$ and the initial *if_score* was set to 1. Equation (3.3) was iteratively run until the difference between the two successive iterations converged to a threshold value, set to 0.0001. To

generate a summary, the sentences from the local document were sorted based on their *if_score* and the top sentences added to the summary until the summary word limit was reached.

3.1.2 Similarity Measures

One important component in the construction of an Affinity Graph is the similarity search to find related documents and sort sentences. Both Wan & Xiao (2010) and Goyal et al. (2013) used the standard Cosine Similarity to measure similarity. The approach was based on a past study of a document expansion network (Tao, Wang, Mei, & Zhai, 2006). The use of Cosine Similarity for document summarization was also discussed by Soe-Tsyr & Jerry (2005) and Qiu & Pang (2008.

In Cosine Similarity, documents are represented as vectors in a large multi-dimensional space, one dimension per unique term in the collection the documents are part of. Measuring the cosine of the angle between the vectors calculates similarity. This similarity is defined in Equation 3.4: the normalized dot product between the two vectors:

$$sim_{doc}(d_i d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\left\|\vec{d}_i\right\| \times \left\|\vec{d}_j\right\|} \qquad (3.4)$$

However, other similarity approaches exist. One ranking function that has a similar way to search for document similarity is Okapi BM25[2] (Robertson, Walker, & Beaulieu, 2000). BM25 has gained popularity due to its strong performance in TREC (Svore & Burges, 2009). BM25 ranks a set of documents based on the frequency of terms that appears in a document and the length of the document.

Another search function that can be used to serve similarity search function is the Indri Language Model, where Strohman, Metzler, Turtle, & Croft (2005) used in the Inquery search engine. The Model consists of two main features: Indri Query Language and Indri Retrieval Model[3]. Both

[2] BM is acronym for Best Match

[3] http://www.lemurproject.org/indri/

features are able to support simple and complex queries (such in our case where the query is the whole document), thus we assumed that Indri Language Model would be able to identify a more diverse selection of documents for our 'nearest neighbour' for our local document.

All three of the similarity measures discussed in this section is available in the Lemur Toolkit (Appendix A).

3.1.3 ROUGE Evaluation

For evaluation, we used the ROUGE (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation) (Lin, 2004b), which provides scores for different evaluation metrics (ROUGE n-gram, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU) as discussed in Chapter 2. ROUGE is commonly used to evaluate the quality of a summary by comparing a generated summary to reference summaries (or gold summaries) by counting the number of overlapping words between them. For our experiments, we will report the scores for ROUGE-1 and ROUGE-2. Lin & Hovy (2003) reported that ROUGE-1 and ROUGE-2 are a reliable score due to its high correlation with human assessment. Lin (2004a) also reported that ROUGE-1 works best in evaluating short summaries, such as for news headlines, and ROUGE-2 is better for single document summarization.

The Recall ROUGE-n is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in \{RefSum\}} \sum_{n-gram_n \in S} Count_{match}(n - gram_n)}{\sum_{S \in \{RefSum\}} \sum_{n-gram_n \in S} Count\,(n - gram_n)} \quad (3.5)$$

where n is the length of the *n*-gram, $gram_n$ and $count_{match}(n - gram_n)$ are the maximum number of n-grams co-occurring in a generated summary and a set of reference summaries ($RefSum$). Here the number of n-grams in the ROUGE-N formula will increase with more reference summaries. Thus, a generated summary that contains words shared by more references is favoured by the ROUGE-N measures. We use the -fA option in ROUGE, which causes the average score of the reference

summaries to be calculated). We also used the "-l 100" in ROUGE to shorten the summaries to 100 words.

For the documents, we work with (DUC2002) two manually generated reference summaries were provided for each document.

3.2 Experiment Setup

The experiments were conducted on the Document Understanding Conference (DUC2002)[4] data set, focussing on Task 1: generate an automatic summary of 100 words or less from a single news document. The organisers of DUC provided 567 English news articles that were manually categorized into 59 groups (e.g. events and biography), and were at least ten sentences long.

For our experiments, stop words were removed, and the Porter stemmer (Porter, 1980) was used to stem the sentences in all of the documents. We used a search engine toolkit from the Lemur Project[5] to calculate the three different approaches to similarity.

3.2.1 Summarisation Settings

We established three types of summarisation settings: (Figure 3.4):

1. **Local Document** uses only information from within the local document itself.
2. **Expanded Documents** use information from the expanded documents only, and
3. **Local+Expanded Documents** use information from both the local document and the expanded documents.

We also defined two document types as input for the expanded document relationship:

- A full document (Full_Doc) is where all sentences in the documents are used, and
- Lead paragraph (Lead_Para), where only the first 100 words of the documents are used.

[4] http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

[5] http://www.lemurproject.org/

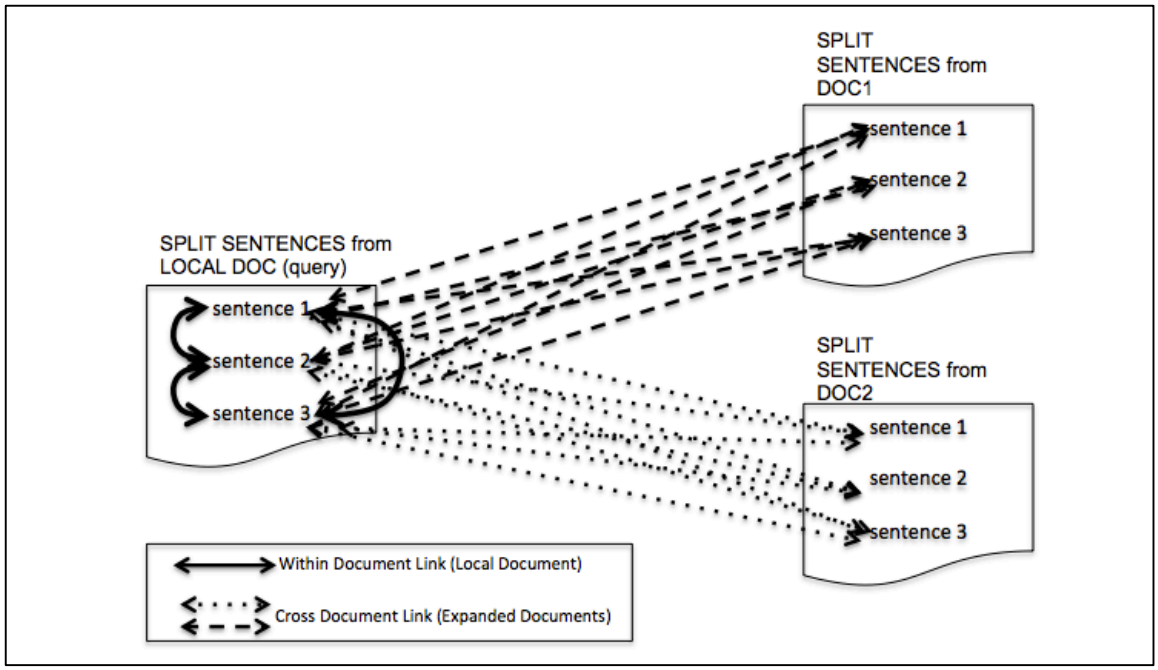


Figure 3.4: Sentence-link relationship

The reason for this different setup is that we wanted to see if the length of the documents has any effect on selecting the sentences from the Local Document. Since the DUC documents are news articles, we believed that the lead paragraph commonly contains a summary of the news report itself (Brandow et al., 1995; Salton et al., 1997). We repeated the experiments using both document types in the Expanded and Local + Expanded neighbourhood information settings.

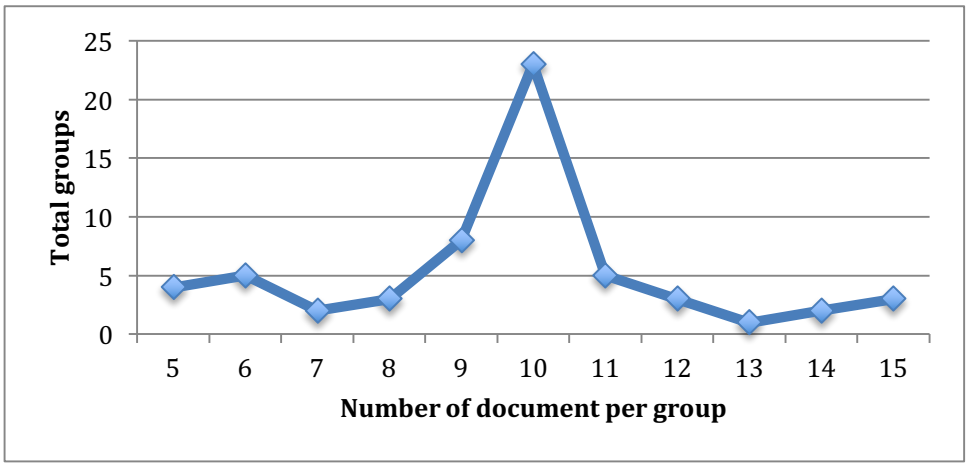


Figure 3.5: Total numbers of documents in groups

For this experiment, we wanted to examine if the 59 manual categories of documents had any effect on the produced summary. We, therefore, conducted experiments using expanded documents limited to one of the different document groupings. Each of the 59 groups has between 5 and 15 documents, see Figure 3.5. To make sure all groups have the same number of documents, we only used four expanded documents to support the local document summaries (one document will be the local document to be summarized).

3.3 Results

This section will discuss the results of our experiments. We first defined the Baseline summary. The organisers of DUC2002 provided a so-called baseline summary, which were the first 100 words of each document. The ROUGE-1 and ROUGE-2 scores of these summaries are provided in Table 3.1.

Table 3.1: Recall ROUGE Score for Lead Paragraph Summary

	ROUGE-1	**ROUGE-2**
BASELINE	**0.471** (0.463 - 0.478)	**0.222** (0.214 - 0.230)

As discussed in Chapter 2, both Wan and Xiao (2010) and Goyal et.al (2013) did not discuss the Baseline summaries provided by DUC and used their own local documents as a baseline. In Table 3.1, the ROUGE score of the lead paragraph was found to give a higher value than the score of summaries from the local documents. Therefore, we decided to use the lead paragraph as our baseline.

We also compared our results with Wan and Xiao (2010) and Goyal et.al (2013) results (see Table 3.2 and Table 3.3). Note that, the parameter for the ROUGE setting in Wan and Xiao (2010) and Goyal et.al (2013) is not known[6], thus, we cannot be certain that we are using the same parameter settings. We tried to match their scores for the DUC systems[7] reported in their paper by testing

[6] We contacted the authors but did not get any response.

[7] The DUC02 baselines and summaries from other participating systems are downloadable from the NIST website (need permission to login).

different parameter settings. At the end, we used the following parameters when invocing the ROUGE script: *-n2 -m -2 4 -u -c95 -r1000 -fA -p0.5 -t0 -l100*. This invocation was the same parameter setting reported in Harman, Steinberger, Poesio, Kabadjov, & Ježek, (2007) and Lloret & Palomar (2010). Even though we did not get the same ROUGE score reported by both papers (for the other DUC systems), we believe that the parameter we used was the best setting to compare with their results.

In Wan and Xiao (2010) and Goyal et.al (2013) the neighbourhood settings were given the following names:

- IntraLink: our Local Document setting
- InterLink: our Expanded Document setting
- UniformLink: our Local+Expanded Document setting

Table 3.2: Recall ROUGE Score for Summary by Wan and Xiao (2010)

	ROUGE-1	**ROUGE-2**
IntraLink	0.460	0.192
UniformLink (k=1)	0.460	0.195
UniformLink (k=5)	0.460	0.195
UniformLink (k=10)	0.464	0.198
InterLink (k=1)	0.460	0.194
InterLink (k=5)	0.464	0.198
InterLink (k=10)	0.463	0.197

Table 3.3: Recall ROUGE Score for Summary by Goyal et.al (2013)

	ROUGE-1	**ROUGE-2**
Intralink	0.450	0.190
PMI	0.452	0.192
MI	0.460	0.200
Bernoulli	0.461	0.202
Uniformlink	0.460	0.199
bern-neB	0.462	0.204
bern+neB	0.464	0.207

Wan and Xiao (2010) and Goyal et.al (2013) reported that their Affinity Graph summaries significantly improved their baseline (Local Document summaries or Intralink). However, they did not discuss and compare their results with the DUC baselines. They also did not discuss if the

different parameters of their settings are significantly better or worse than one another (except for Wan & Xiao (2010) where they only discussed the different number of related documents). Based on the results in Table 3.2 and 3.3, neither of the summarization systems produced by Wan & Xiao (2010) or Goyal et al. (2013) able to improve upon the baseline summaries that we produced in Table 3.1.

We noticed that the results in both tables are similar (and perhaps did not show significant different between each other). However, it is important to discuss their results in this thesis, to show that that our Affinity Graph algorithm does improve the summaries compared with its Local Document (or IntraLink) summaries, as reported by Wan and Xiao (2010) and Goyal et.al (2013) in their work. Note that originally, the aim of this work is to improve Local Document summaries, thus it is critical to discuss similar work by others.

3.3.1 Summaries with Affinity Graph Algorithm

The first thing that we were interested to understand was how the choice of the similarity measure affected summary accuracy. We compared the Cosine, Okapi BM25, and Indri Language Model measures using (k=10) expanded documents, as in Table 3.4. We found that Cosine Similarity generally resulted in the best accuracy for both ROUGE-1 and ROUGE-2. However, BM25 gave the best overall ROUGE-1 score using just the lead paragraph of expanded documents. Using the Indri Language Model did not result in high accuracy.

The summaries supported by the expanded documents using the Indri Language Model also did not result in summaries that were more accurate than Local Document summaries, unlike the other two similarity measures. Further investigation on the effects of using the Indri Language Model will be discussed on Section 3.4.

Table 3.4: ROUGE Score for Cosine Similarity, Okapi BM25 and Indri Language Model (k=10)

	COSINE SIMILARITY		OKAPI BM25		INDRI LANGUAGE MODEL	
	ROUGE-1	**ROUGE-2**	**ROUGE-1**	**ROUGE-2**	**ROUGE-1**	**ROUGE-2**
Local Document	0.429 (0.422 - 0.436)	0.17 (0.159 - 0.175)	0.414* (0.406 - 0.422)	0.164 (0.156 - 0.172)	**0.396*** (0.388 - 0.403)	**0.141** (0.133 - 0.149)
Expanded	0.436# (0.429 - 0.444)	0.175# (0.167 - 0.184)	0.418#+ (0.409 - 0.425)	0.164 (0.156 - 0.173)	0.382#+ (0.377 - 0.392)	0.133# (0.133 - 0.149)
Local+Exp	0.438# (0.430 - 0.444)	0.176# (0.168 - 0.184)	0.417^ (0.410 - 0.426)	0.165# (0.156 - 0.173)	0.384#^ (0.374 - 0.389)	0.131# (0.123 - 0.138)
Expanded Lead_Para	**0.439#** (0.431 - 0.446)	**0.176#** (0.169 - 0.185)	**0.441#** (0.434 - 0.449)	**0.185#** (0.176 - 0.193)	0.377#~ (0.370 - 0.385)	0.126# (0.119 - 0.134)

 # statistically significant (p-value < 0.05) compared to Local Document summaries (from the same similarity measures in ROUGE-1 and ROUGE-2)

** statistically significant (p-value < 0.05) compared to Cosine Local Document summaries*

+ statistically significant (p-value < 0.05) compared to Cosine Expanded Document summaries

^ statistically significant (p-value < 0.05) compared to Cosine Local+Expanded Document summaries

~ statistically significant (p-value < 0.05) compared to Cosine Expanded Lead_Para summaries

We performed paired t-test between all settings. There was significant improvement compared to Local Document summaries for all similarity measures except for the Local+Expanded Document settings in Okapi BM25. A paired t-test between Cosine Similarity and Okapi BM25 for Expanded Lead Paragraph showed no significance difference. The same results can be seen in the ROUGE-2 scores, where all settings (except for Expanded Document in Okapi BM25) showed significant improvement over the equivalent setting for Local Document. Overall we can see that both Cosine and Okapi BM25 performed well, but neither shows overall superiority. The choice of Cosine Similarity by Wan and Xiao (2010) and Goyal et.al (2013) is supported by our work. No other measure was found to be superior.

The next experiment was developed to measure the effect of using different numbers of *k* documents for expansion: *k*=1, 2, 3, 4, 5, 15, and 20. We were also interested to find if by increasing the number of expanded documents and expanded lead paragraphs, would improve the accuracy to support single document summaries.

Figure 3.6 shows the results. No difference is seen in the pattern for Expanded, Expanded Lead_Para, and Local+Expanded [8] documents. Both Expanded Lead_Para and Local+Expanded documents show the highest score when *k*=5. A different pattern is shown for Local+Lead_Para, where the score is much lower than the other settings, however it increases as more documents are used.

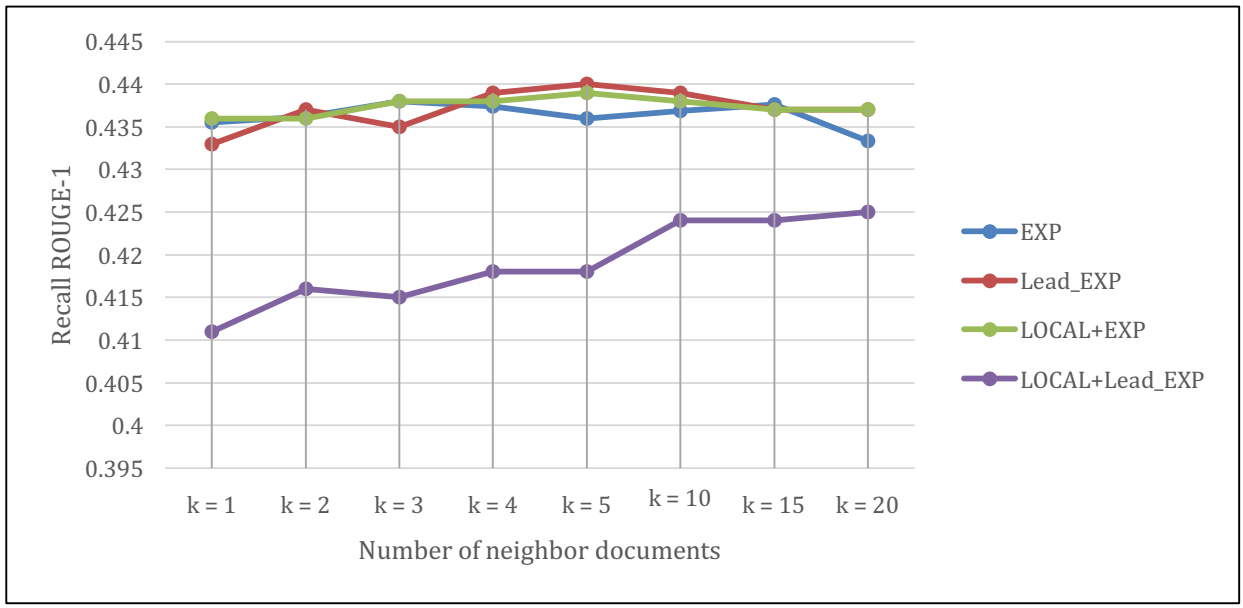


Figure 3.6: Recall ROUGE-1 score for Expanded Document (EXP), Expanded Lead Paragraph (Lead_Para), Local+Expanded Document (LOCAL+EXP) and Local+Lead_Para (LOCAL+Lead_Para) with different values of k

We tested for significance for ROUGE-1 for all settings, and in Expanded Documents, we found that there was no difference from *k=1* to *k=15*. It showed that only the lowest score (*k*=20) gave significance difference when compared to the highest ROUGE-1 score (*k*=3). For Expanded Lead_Para, there were no significant difference between *k*=4 to *k*=20 for the ROUGE-1 scores, when compared with its highest ROUGE-1 score, which is when *k*=5. This shows that the ROUGE scores

[8] For this experiment and the LOCAL+Lead_Para, the Local Document is included in the expanded documents set. Thus, the total documents used in this experiment are *k +1* (where *k* represents the Expanded Document and *+1* represent the Local Document).

for Expanded Document and Expanded Lead_Para did not show constant improvement when more documents were used.

For the LOCAL+EXP documents, only the lowest ROUGE-1 scores (from k=2) showed a significance difference in the Recall ROUGE-1 when compared with k=5 (the highest ROUGE-1 score). In the Local+Expanded using lead paragraph (LOCAL+Lead_Para), we see that k=20 gave the best results, but it only showed significance when compared to small values of k (from k=1 to k=5).

It appears that in Figure 3.6, good results can be obtained when using a modest number of appropriately chosen full documents for the expansion, with k=3 to k=5 yielding results all within 0.1 percentage points of each other. All of the results from the full document settings were also significantly better when compared with the lead paragraph with the same k values.

For our next experiments, we explored the use of manually marked up categories to constrain the set of documents to be expanded from. This setup is to test the scenario such that if we have manually grouped/clustered documents, can the grouping be exploited in the summary generation? This neighbourhood setting is assumed to have just enough information for the Local Document, since all the documents were grouped by DUC to represent a certain event and categories.

We used the same document neighbourhood information setting (the Expanded Documents and Local+Expanded Documents) for the group dataset experiments. As described earlier, we limited the number of Expanded Documents (k) from 1 to 4 documents only, where the 5th document is the Local Document to be summarized. For the groups with more than 5 documents, we took the 4 documents with the highest similarity value. This was to make sure all local documents had the same value of k for the experiments (this also applied to Local+Expanded document settings). The reason for this, is that we wanted to make sure all 59 groups were included in the experiment. We also used the lead paragraph documents (Lead_Para) in the group dataset experiments.

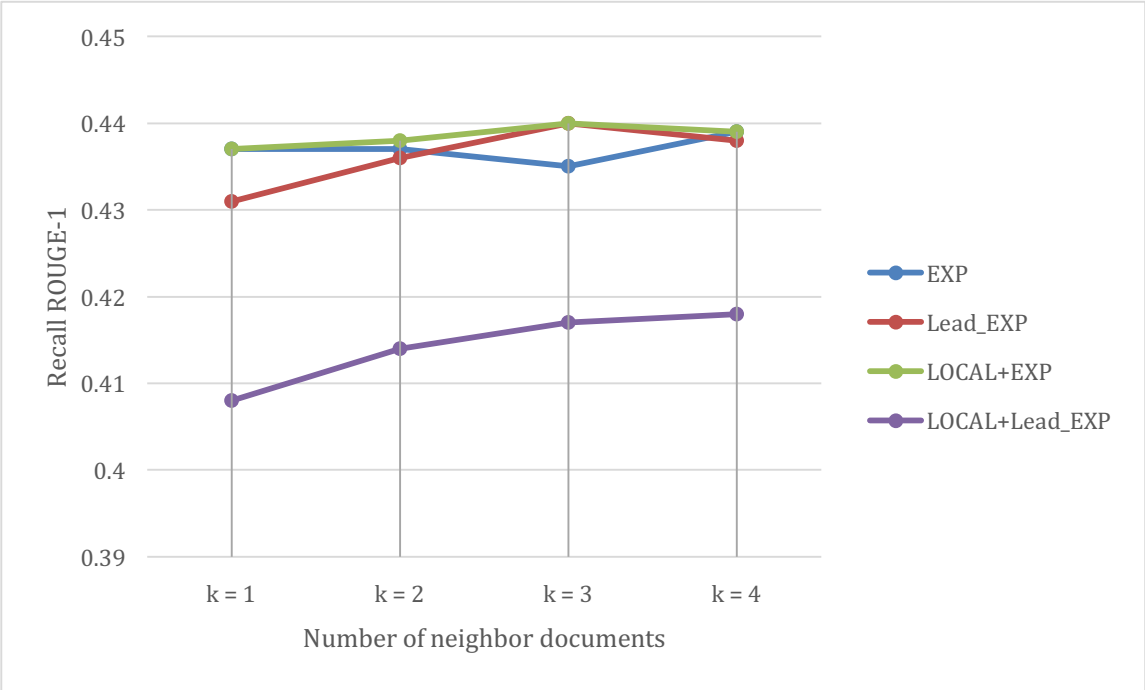


Figure 3.7: Recall ROUGE-1 score for Expanded Document (EXP), Expanded Lead Paragraph (Lead_Para), Local+Expanded Document (LOCAL+EXP) and Local+Lead_Para (LOCAL+Lead_Para) with different values of k in a group dataset setting

The results in Figure 3.7 showed the same pattern as in Figure 3.6, where Expanded, Expanded Lead_Para and Local+Expanded documents give better scores than the LOCAL+Lead_Para documents. The significance test for the Recall ROUGE-1 scores between the full document and lead paragraph also showed a statistically significance difference only when compared with fewer documents used as the expanded document. Based on this result, we believed that having limited information for the expanded documents might not provide the relevant information that we need to improve single document summaries.

Note that the results for manually grouped documents in a dataset were never quite as good as those for which the full documents in a dataset were used for expansion. We expect that this is because more information is available when all documents in the dataset are considered to be the nearest neighbour documents. Thus, having as many documents to be chosen from, might give a better support for the single document summarization system. We might consider repeat the

experiments using groups with $k>4$. However, the ROUGE scores have similar pattern with the summaries using all documents in the dataset. Thus, we assumed that the results might not show any statistical significance difference.

3.4 Summary Evaluation

In order to have a better understanding of our auto-created summaries, we took a closer look at the documents, gold standard summaries, and the automated summaries. We randomly selected a document with 10 sentences (short length document), 45 sentences (medium length document) and 103 sentences (long length document) together with its summaries.

Short Length Document

Figure 3.8 showed the original document and its two gold standard summaries (human created summary) for the document AP900128-0063. The highlighted part of the document is the Baseline Summary (Lead Paragraph) of the document as provided by DUC02. Both gold standard summaries were used as the model summary in ROUGE and were used to compare with the generated summaries. The Figure 3.8 also showed an example of the summaries generated using the three similarity measures (for Expanded Documents with $k=10$).

The first part of our experiments was to explore the similarity measures to be used for searching related document. Cosine similarity was not beaten by Okapi BM25 and Language Modelling. Okapi BM25 tends to over penalize long documents (Lv & Zhai, 2011); thus, having a full document as a query itself may not be the ideal way to obtain an *if_score*. However, using the lead paragraph as the query for document similarity in Okapi BM25, proved to be more successful to generate the summaries as in Table 3.2. Okapi BM25 outperformed Cosine Similarity and Language Model with the average ROUGE-1 score of 0.44.

ORIGINAL DOCUMENT (AP900128-0063) TITLE: San Francisco Routs Broncos in Super Bowl	
The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade and tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries. San Francisco won the National Football League championship game in 1989, 1985 and 1982. The Broncos, on the other hand, lost the last four Super Bowl games they have played. San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions. He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever. Montana left the game with nearly 11 minutes to play. In four Super Bowls, he has thrown 11 touchdowns and no interceptions. For Denver quarterback John Elway, it was a day of futility, ending with his third Super Bowl defeat. He missed eight of his first 10 passes and was intercepted twice and fumbled once. By halftime the score was 27-3. With their third loss in four years, the Broncos have now been outscored 136-40. San Francisco was boringly perfect, doing more than even the experts who made them favorites by nearly two touchdowns after a 14-2 season and a waltz through the playoffs.	
GOLD STANDARD	
Abstract Summary 1	**Abstract Summary 2**
The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. This was the 49ers fourth win in four tries, tying them with the Pittsburgh Steelers for the NFL championship. San Francisco quarterback Joe Montana set several Super Bowl records; five touchdown passes; 13 straight pass completions; a third Super Bowl MVP award. San Francisco's point total was the most ever. In four Super Bowls, Montana has thrown 11 touchdowns and no interceptions. It was a day of futility for Broncos' quarterback John Elway as his team suffered its third Super Bowl defeat.	The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. The 24th Super Bowl also generated other statistics. The 49ers, having also won in 1989, 1985 and 1982, tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries. San Francisco quarterback Joe Montana, broke the record for touchdown passes, 5. He also set a record for straight pass completions, 13. Montana set five Super Bowl career records, including his third Super Bowl MVP award, and San Francisco's point total was the most ever.
SUMMARY 1: COSINE SIMILARY	
Full document summary	**Lead paragraph summary**
[1] San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions. [2] San Francisco won the National Football League championship game in 1989, 1985 and 1982. [3] The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. [4] The Broncos, on the other hand, lost the last four Super Bowl games they have played. [5] The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion.	[1] The Broncos, on the other hand, lost the last four Super Bowl games they have played. [2] San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions. [3] The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. [4] The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade and tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection.
SUMMARY 2: OKAPI BM25	
Full document summary	**Lead paragraph summary**
[1] San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions. [2] For Denver quarterback John Elway, it was a day of futility, ending with his third Super Bowl defeat. [3] He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever. [4] In four Super Bowls, he has thrown 11 touchdowns and no interceptions. [5] He missed eight of his first 10 passes.	[1] San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions. [2] He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever. [3] The San Francisco 49ers routed the Denver Broncos 55-10 Sunday in the most lopsided Super Bowl victory ever. [4] For Denver quarterback John Elway, it was a day of futility, ending with his third Super Bowl defeat.
SUMMARY 3: LANGUAGE MODEL	
Full document summary	**Lead paragraph summary**
[1] San Francisco was boringly perfect, doing more than even the experts who made them favorites by nearly two touchdowns after a 14-2 season and a waltz through the playoffs. [2] With their third loss in four years, the Broncos have now been outscored 136-40. [3] The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade and tied the Pittsburgh Steelers as a pinnacle of Super Bowl perfection with four wins in four tries. [4] San Francisco won the National Football League championship game in 1989, 1985 and 1982.	[1] San Francisco was boringly perfect, doing more than even the experts who made them favorites by nearly two touchdowns after a 14-2 season and a waltz through the playoffs. [2] With their third loss in four years, the Broncos have now been outscored 136-40. [3] He also set five Super Bowl career records, including his third Super Bowl Most Valuable Player award and San Francisco's point total was the most ever. [4] Montana left the game with nearly 11 minutes to play. [5] The 49ers' win in the 24th Super Bowl made them the first repeat NFL champion in a decade.

Figure 3.8: Short-length document summaries generated using Cosine Similarity, Okapi BM25 and Language Model (using Expanded Full Document and Lead Paragraph)

Looking further, we can see that all three summaries generated by different similarity measures have different topics extracted from the local document. In the abstract summaries created manually by DUC, we can see the document is summarised into two main topics; the Super Bowl team (*The San Francisco 49ers/Denver Broncos*) and the players (*Joe Montana/John Elway*). Both topics were mentioned in the baseline summary, where this information is in the first four sentences. Thus we can see that for short-length documents, the main topics of the documents are available in the first few sentences of the document.

In the Affinity Graph summaries, we can see differences between the summaries created by Cosine and Okapi BM25. In Cosine, the topic of the summaries was on the Super Bowl team. Both summaries (full document and lead paragraph) contain information regarding the teams, except for one sentence where it extracts the sentence on the player that made a touchdown record.

However, for the summaries generated using the Okapi BM25, most of the sentences extracted by the Affinity Graph summarizer were on the Super Bowl players. It was interesting to see that different similarity measure methods are able to extract different topic for the summaries. Even though the sentences were not in the same order as the documents, it is still understandable.

But all four summaries by Cosine Similarity and Okapi BM25, extracted a common sentence ("*San Francisco quarterback Joe Montana made five touchdown passes, three to Jerry Rice, breaking a Super Bowl record for touchdown passes on a day on which he also set a record with 13 straight pass completions*"), which we agree that would be the main information for the documents. This information is also mentioned in the abstract summaries created by human experts.

We can see that the Indri Language Model failed to capture the meaning of the document. This could be the reason that the non-related documents and/or sentences may be selected, and gave a great effect on why the Indri Language Model performed the worst in all similarity measures. The summaries generated have a mix of topics and the sentence order was not as good as the other summaries.

Based from this, we agreed that the summaries generated by Cosine Similarity and Okapi BM25 for short length documents are equally good and able to give different views of the topic from the document. This does not show in the ROUGE score, where both similarity measures were not significantly better between each other. The difference of ROUGE score for both similarity measures was very small.

Medium Length Document

We randomly selected document FT933-10917, which has 45 sentences. Figure 3.9 shows the documents with baseline (highlighted), its gold standard summaries and Affinity Graph generated summaries (the full text can be found in Appendix B). We can see in the baseline summary, it only contained a short description on John Major's characteristics and his interview with Michael Brunson.

Both abstract summaries discussed John Major's political image, the ITN interview with Michael Brunson and the Christchurch by-election. For the Affinity Graph summaries, the Cosine measure resulted in the only summary that extracted the first sentence ("*THE revelation that John Major is capable of candid, blunt and salty language when …*") and the ("*What piffle.*"), which was included in the baseline summary. For the other sentences resulting from Cosine Similarity, it extracted information on the "interview", with the Lead Paragraph method extracting the second sentence, where the ITN's interview was mentioned.

For the Okapi BM25 summaries, Full document and Lead Paragraph's summary extracted 2/3 of the same sentences. As for the Language Model summaries, again it extracted the least informative sentences (compared to the other summaries – including the abstract and the baseline summaries). We can see that the Language Model summaries extracted sentences with the most number of proper nouns (such as *Olivier Blanchard, Rudiger Dornbusch, Stanley Fischer, Franco Modigliani, Paul A Samuelson* and *Robert Solow*).

ORIGINAL DOCUMENT (FT933-10917)
TITLE: Hawks & Handsaws: A few blunt words

THE revelation that John Major is capable of candid, blunt and salty language when talking off-the- record to friendly journalists has surprised some people.
It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good.
With luck, it is reckoned, Major's image as a leaden-tongued wimp may undergo correction.
What piffle.
Major is a gonner, especially after this week's revolt of the wooden-tops in the Christchurch by-election, where a Conservative majority of 23,015 at last year's general election was converted into a 16,427 majority for the Liberal Democrats.
Fifteen months too late, the voters of Christchurch rounded on the Tories with a malignant and squeaky fury.
In reality, all politicians, not just Major, are far more candid and salty when chatting in private than when speaking in public.
In public, they have to be careful of what they say, so their utterances achieve a horrible mattness.
But in private they relax.
Their syntax disappears.

GOLD STANDARD

Abstract Summary 1	Abstract Summary 2
John Major's public image as a wimp may have changed following a candid interview with Michael Brunson, ITN's political editor. Major used a variety of epitaphs to describe certain colleagues who are not his favorites. It is a change that comes too late. The Conservative majority in last year's general election in Christchurch turned into a victory for the Liberal Democrats in this year's by-election. In reality, the language of all politicians, including Major, changes when they speak in private. Gone are the crafted sentences as a certain saltiness creeps in and the politicians swear and joke.	Some believe that the recording of John Major's conversation with Michael Brunson, ITN's political editor, in which Major used salty language will improve Major's image as a wimp. Not so. Major is gone, especially after the Christchurch by-election where a Conservative majority in last year's general election was transformed to a liberal Democrats majority of 16,427. Actually, all politicians are far more candid when speaking privately. To demonstrate, I spoke yesterday with both Major and John Smith, Labor Party leader. Using a scrambler, to guarantee privacy, I provoked some frank discussion. The rambling responses of both were liberally sprinkled with expletives.

SUMMARY 1: COSINE SIMILARY

Full document summary	Lead paragraph summary
[1] I can live with that, though why the artsy-fartsies should receive any dispensation is a puzzle. [2] What piffle. [3] It really is a spectacle. [4] To show you what I mean, I spoke yesterday to John Major and John Smith. [5] I told him I had been impressed with his interview with Andrew Marr in The Independent on Thursday, in which he sharpened up his promise to introduce meaty political reforms (if he ever gets elected), including a referendum on proportional representation. [6] First, I tackled Major. [7] THE revelation that John Major is capable of candid, blunt and salty language when …	[1] What piffle. [2] I can live with that, though why the artsy-fartsies should receive any dispensation is a puzzle. [3] It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good. [4] Unfairly or not, you are drawing the blame for all life's unpleasantnesses, let alone the cock-ups'. [5] THE revelation that John Major is capable of candid, blunt and salty language when talking off-the- record to friendly journalists has ...

SUMMARY 2: OKAPI BM25

Full document summary	Lead paragraph summary
[1] I told him I had been impressed with his interview with Andrew Marr in The Independent on Thursday, in which he sharpened up his promise to introduce meaty political reforms (if he ever gets elected), including a referendum on proportional representation. [2] It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good. [3] Major is a gonner, especially after this week's revolt of the wooden-tops in the Christchurch by-election, …	[1] It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good. [2] Likewise with political and constitutional reform, Michael, for by the L - - d, tho' I should beg wi'lyart pow, I'll laugh, an' sing, an' shake my leg, as lang's I dow]' After that, I thought of telephoning Wing-Commander Paddy Ashdown, leader of the Liberal Democrats, to solicit his views on Christchurch. [3] Major is a gonner, especially …

SUMMARY 3: LANGUAGE MODEL

Full document summary	Lead paragraph summary
[1] I said: 'Did you read, John, what Olivier Blanchard, Rudiger Dornbusch, Stanley Fischer, Franco Modigliani, Paul A Samuelson and Robert Solow wrote, in just one article, in the FT this week? [2] Likewise with political and constitutional reform, Michael, for by the L - - d, tho' I should beg wi'lyart pow, I'll laugh, an' sing, an' shake my leg, as lang's I dow]' After that, I thought of telephoning Wing-Commander Paddy Ashdown, leader of the Liberal Democrats, to solicit his views on Christchurch. [3] But I have left out the swear-words because the new Financial Times Style Guide states that . . .	[1] I said: 'Did you read, John, what Olivier Blanchard, Rudiger Dornbusch, Stanley Fischer, Franco Modigliani, Paul A Samuelson and Robert Solow wrote, in just one article, in the FT this week? [2] But I have left out the swear-words because the new Financial Times Style Guide states that 'the gratuitous use of expletives or obscenities is discouraged . . . Four-letter expletives will usually be confined to infrequent use in the review (Arts) pages'. [3] 'I mean . . . how did it come about, Michael . . . like, Christchurch, y'know - load of . . . let me put it to …

Figure 3.9: Medium-length document summaries generated using Cosine Similarity, Okapi BM25 and Language Model (using Expanded Full Document and Lead Paragraph)

For medium length document, Cosine Similarity generated a better summary compared to the Okapi BM25 and the Language Model. It appears that Cosine Similarity was able to extract sentences that contained fewer proper nouns or spoken sentences; thus, the summary makes more sense. However, distinct differences can be seen for both summaries; unlike in the short length summaries where Cosine Similarity and Okapi BM25 extracted the same sentences.

Long Length Document

Figure 3.10 shows the summaries for LA101590-0066, and we considered this as a long-length document (103 sentences – full document in Appendix C). In the abstract summaries, it included information on the birth, life and death of Leonard Bernstein, as well as his work (as the conductor of the New York Philharmonic) and the success of the "West Side Story". However, in the baseline summary, only the news of death of Bernstein was mentioned.

Again, we can see that all three similarity measures generated different summaries (but have at least one same sentence for its Full Document-Lead Paragraph pair). For the Cosine Similarity, it focused on his work as a music conductor, with the mention of the *New York Philharmonics*. Both of the Okapi BM25's summaries extracted different summaries, where its Lead Paragraph summaries extracted longer sentences and the "*West Side Story*" was mentioned only in Okapi BM25's full document summaries. And again, the Language Model's summaries extracted a different part of the document, and very different from the abstract summaries and we believed that is the reason Language Model have the lowest ROUGE-1 score compared to the others.

For the long documents, we can see varieties of summaries for different similarity measures, but none of the auto-generated summaries mentioned the birth and death of Leonard Bernstein; the summaries only focused on his career. However, Cosine Similarity's summaries gave a more relevant history of Bernstein's career, compared to the other similarity measures.

<table>
<tr><th colspan="2">ORIGINAL DOCUMENT (LA101590-0066)
TITLE: Leonard Bernstein Dies; Conductor, Composer; Music: Renaissance Man of His Art Was 72. The Longtime Leader Of The N.Y. Philharmonic Carved A Niche In History With 'West Side Story.</th></tr>
<tr><td colspan="2">Leonard Bernstein, the Renaissance man of music who excelled as pianist, composer, conductor and teacher and was, as well, the flamboyant ringmaster of his own nonstop circus, died Sunday in his Manhattan apartment. He was 72 .
Bernstein, known and beloved by the world as "Lenny," died at 6:15 p.m. in the presence of his son, Alexander, and physician, Kevin M. Cahill, who said the cause of death was complications of progressive lung failure.
On Cahill's advice, the conductor had announced Tuesday that he would retire.
Cahill said progressive emphysema complicated by a pleural tumor and a series of lung infections had left Bernstein too weak to continue working .
In recent months, Bernstein canceled performances with increasing frequency. His last conducting appearance was at Tanglewood, Mass., on Aug. 19.
Bernstein was the first American-born conductor to lead a major symphony orchestra, often joining his New York Philharmonic in playing his own pieces, while conducting from the piano.
He etched other niches in history by composing the indelible "West Side Story" and teaching a generation about classical music via the innovative television series "Omnibus."
Exhibiting remarkable talent and expertise in four areas that most artists wish they possessed in merely one, Bernstein still might have remained an obscure musician without the unique theatrical flair that dominated his personal as well as professional life.</td></tr>
<tr><th colspan="2">GOLD STANDARD</th></tr>
<tr><th>Abstract Summary 1</th><th>Abstract Summary 2</th></tr>
<tr><td>Leonard Bernstein, the Renaissance man, died Sunday at 72 from lung disease. He had remarkable talent and expertise in conducting, composing, playing the piano, and teaching, which he combined with a unique theatrical flair in both his professional and private lives. He was born of Russian Jewish immigrants, began the piano at age 10, and was educated at Boston Latin School and Harvard University. He was mentored by the great musicians of the era. Perhaps his greatest successes were conducting the New York Philharmonic and composing "West Side Story" for Broadway. He was a heavy smoker and drinker with an uproarious, liberal, life style.</td><td>Leonard Bernstein, the flamboyant Renaissance man of music who excelled as pianist, composer, conductor and teacher died Sunday in his Manhattan apartment. He was 72 and had suffered from progressive lung failure. Mr. Bernstein, the internationally acclaimed conductor of the New York Philharmonic from 1957 until 1968, was the first American-born conductor to lead a major symphony orchestra. Harvard educated and mentored by several musical giants, including Aaron Copeland, he brought classical music to the masses via his innovative television series "Omnibus," and made an indelible mark on American popular music with his composition, "West Side Story".</td></tr>
<tr><th colspan="2">SUMMARY 1: COSINE SIMILARY</th></tr>
<tr><th>Full document summary</th><th>Lead paragraph summary</th></tr>
<tr><td>[1] Successful as a pianist, composer and conductor, Bernstein, according to Joan Peyser in a controversial biography, consulted psychiatrists because of his internal conflict over the three pursuits.
[2] "It is impossible for me to make an exclusive choice among the various activities," Bernstein wrote in 1946.
[3] With no rehearsal, a hangover and three hours sleep, Bernstein was to conduct a complex program broadcast nationwide on CBS radio.
[4] Bernstein was the first American-born conductor to lead a major symphony orchestra, often joining his New York Philharmonic in playing his own pieces, while conducting from the piano.
[5] He left the orchestra …</td><td>[1] Bernstein was not to get his own orchestra until he took over the New York Philharmonic in 1957-58.
[2] Bernstein was the first American-born conductor to lead a major symphony orchestra, often joining his New York Philharmonic in playing his own pieces, while conducting from the piano.
[3] He left the orchestra in 1969, after a record 11-year tenure at the helm, to have more time for composing and guest conducting.
[4] "Some conductors mellow with age," commented Times music critic Martin Bernheimer when Bernstein conducted the Los Angeles Philharmonic at UCLA in 1986.
[5] "The influence of Mitropoulos on my life, …</td></tr>
<tr><th colspan="2">SUMMARY 2: OKAPI BM25</th></tr>
<tr><th>Full document summary</th><th>Lead paragraph summary</th></tr>
<tr><td>[1] His best and best-remembered work, "West Side Story," debuted in 1957.
[2] "He had no children of his own and I had a father whom I loved very much but who was not for this musical thing at all.
[3] "An assessment of Bernstein must include his talent and contribution as a teacher and popularizer of music, a role that has set him apart most from other performers," conductor, historian and Bard College President Leon Botstein wrote in Harper's in 1983.
[4] "I have gone through all the conductors I know of in my mind and I finally asked God whom I ….</td><td>[1] "An assessment of Bernstein must include his talent and contribution as a teacher and popularizer of music, a role that has set him apart most from other performers," conductor, historian and Bard College President Leon Botstein wrote in Harper's in 1983.
[2] "The influence of Mitropoulos on my life, on my conducting life is enormous and usually greatly underrated or not known at all," Bernstein wrote years later, after his mentors had all died, "because ordinarily the two great conductors with whom I studied are the ones who receive the credit for whatever conducting prowess I have, namely Serge Koussevitzky and …</td></tr>
<tr><th colspan="2">SUMMARY 3: LANGUAGE MODEL</th></tr>
<tr><th>Full document summary</th><th>Lead paragraph summary</th></tr>
<tr><td>[1] Exhibiting remarkable talent and expertise in four areas that most artists wish they possessed in merely one, Bernstein still might have remained an obscure musician without the unique theatrical flair that dominated his personal as well as professional life.
[2] "(But) Bernstein, at 68, remains a frenetic combination of orbiting rocket, aerobics master, super-juggler, matinee idol, booming cannon, hysterical mime, heart-rending tragedian, bouncing ball, sky writer, riveting machine, mawkish sentimentalist and danseur ignoble".
[3] When his kindergarten teacher asked "Louis Bernstein" to stand up, he remained seated and looked around the room to see who shared his last name.
[4] Bernstein's programs, Botstein …</td><td>[1] "(But) Bernstein, at 68, remains a frenetic combination of orbiting rocket, aerobics master, super-juggler, matinee idol, booming cannon, hysterical mime, heart-rending tragedian, bouncing ball, sky writer, riveting machine, mawkish sentimentalist and danseur ignoble".
[2] Louis Bernstein (so-named because his maternal grandmother insisted) was born Aug. 25, 1918, in Lawrence, Mass., to two Russian Jewish immigrants.
[3] Exhibiting remarkable talent and expertise in four areas that most artists wish they possessed in merely one, Bernstein still might have remained an obscure musician without the unique theatrical flair that dominated his personal as well as professional life.
[4] Describing the conductor in the same …</td></tr>
</table>

Figure 3.10: Long-length document summaries generated using Cosine Similarity, Okapi BM25 and Language Model (using Expanded Full Document and Lead Paragraph)

Based on the sentence extraction analysis, we can see that Affinity Graph algorithm extracted sentences that contains information in the abstract summaries. There are few misses on the relevant information, especially on the longer documents. We believed that for longer documents, few topics dominated the content of documents, thus it might not have captured the overall information as good as the human-abstract summaries. However, in comparison with the Baseline Summaries, the Affinity Graph summaries are able to capture more information in medium and longer documents. In medium-length document (Figure 3.09), we can see that the Cosine Similarity summary extracted two sentences from the baseline summary, and Okapi BM25 extracted one sentence from the baseline summary. This showed that the Affinity Graph summaries are able to identify the more important sentences that located in the first few part of the document; as the full document is a 'complex' document, where it contained proper nouns, conversations and spoken word (e.g.: *y'know*, *Gie me o'wit an*).

In Figure 3.10, we can see that there are three main topics in the human-abstract summaries, however only one of the topics (the death of Leonard Bernstein) was mentioned in the baseline summary (but the topic was not mentioned in any of the Affinity Graph summaries). Again, we can see that Cosine Similarity and Okapi BM25 extracted more relevant sentences (from one or more topics), where these topics were included in the human-abstract summaries.

In Figure 3.11, we can see that there are only four documents with more than 80 sentences in the DUC2002 dataset. We believed that this may contributed to the low ROUGE scores in Affinity Graph summaries, where 33% of the document have less than 20 sentences, but only 6% of the document has more than 60 sentences.

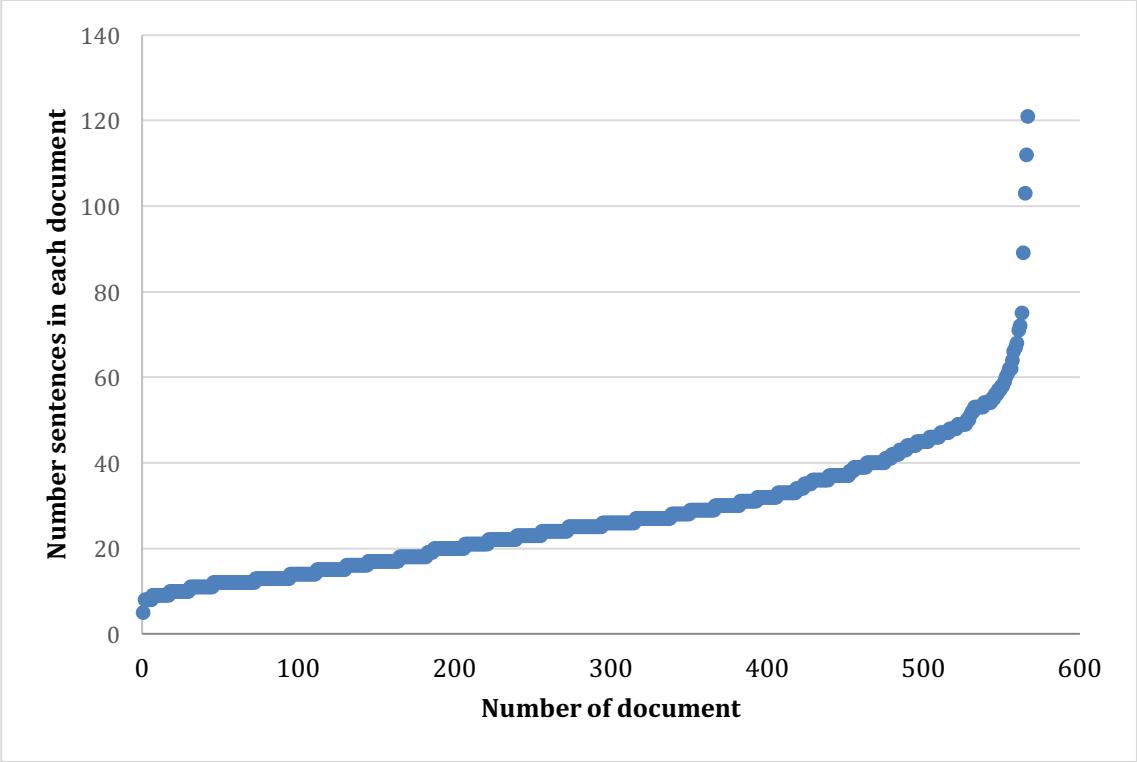


Figure 3.11 Number of sentences for 576 documents in DUC02 dataset

Based on the discussion on in the short-length summaries, where the information in the baseline summary have a high possibility to be included in the human-abstract summaries. Thus, the baseline summary would be able to be very similar with the gold standard (human summaries), and this would have contributed to a high ROUGE score. A more detailed discussion on the ROUGE score correlation is discussed in the next section.

3.4.1 ROUGE Score Correlation

We analysed the correlation between the number of sentences and the Recall ROUGE-1 scores for each of the summaries from the document for the different settings. We wanted to see if the document length has any effect on the ROUGE scores since the DUC dataset has a wide range of document length (the least number of sentences in a document is 5 and the most number is 120). We believed that for documents with more sentences, a summarization system would have a more diverse sentence

selection, thus there is a possibility that irrelevant sentences are selected. A generated summary would have better ROUGE scores when document length is small due to the fact that a summarizer would be able to choose better sentences. To the best of our knowledge, very little past study has been done to analyse the correlation between the document length and ROUGE score.

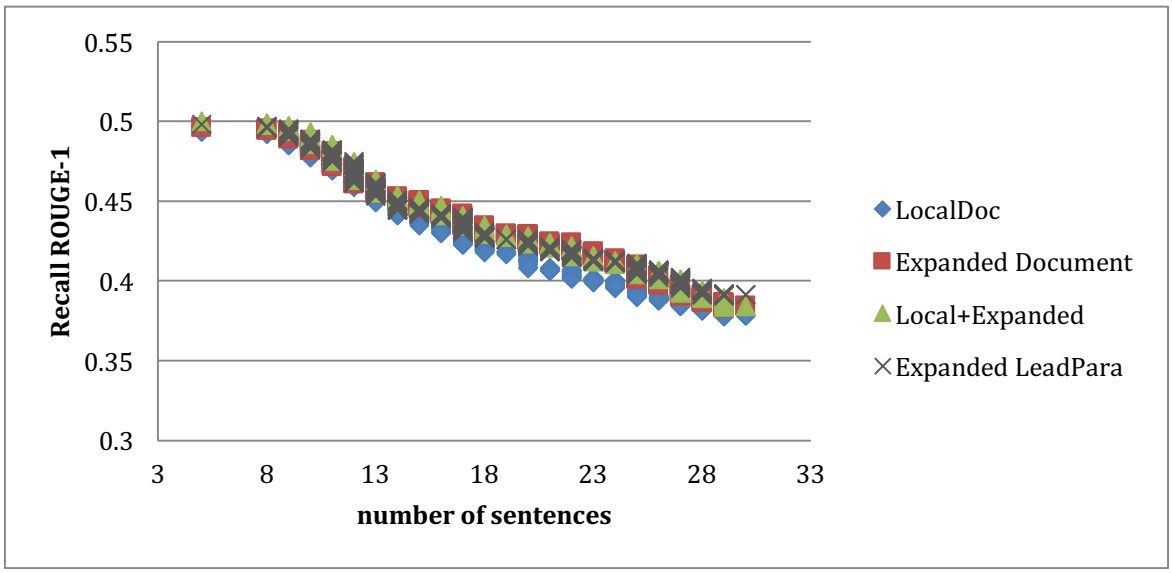


Figure 3.12 Recall ROUGE-1 scores vs. number of sentences (in rolling average)

For our analysis, we sorted the documents according to the number of sentences in descending order. For the ROUGE-1 scores (Local Document, Expanded Documents), we computed the rolling average, where each subset consists of 200 elements. We found that the number of sentence and Recall ROUGE-1 scores have a strong negative correlation ($r=-0.9$, $p < 2.2e-16$) for all settings (Figure 3.12). This shows that summaries from longer documents generally have lower ROUGE scores.

Figure 3.12 also showed that the Local Document setting summaries gave the lowest Recall ROUGE-1 for all documents in the rolling average dataset. This also provides support to our hypothesis that summaries that are supported by their expanded documents would improve the local document summaries.

3.5 Discussion

We recreated (Wan & Xiao, 2010) and (Goyal et al., 2013) experiments, but we failed to reproduce the same results or effectiveness as reported in their paper. We assumed that this is due to two reasons:

1. The ROUGE parameter settings. Both papers did not mention their ROUGE settings, thus we were not sure if we used the same settings for our experiments.
2. The pre-processing of the DUC dataset. We used the split sentences of the DUC02 documents which are downloadable from the DUC dataset (split sentence tool was also provided). However, we further analysed the dataset and discovered that some documents still need to be split. Further pre-processing was conducted to ensure the sentences in the documents were correctly split. Thus, we were not sure if the same pre-processing was done by the previous two studies.

For our experiments, we tested different values for the Affinity Graph parameter settings. The first setting that we tested was the similarity measure. We found that Okapi BM25 was the best similarity measure when only the lead paragraph of expanded documents was considered; however, Cosine Similarity, proved to outperform the others. Cosine Similarity was also proven to be successful in automatic hyperlink generation in a work by Salton et al. (1997). Our discussion in Section 3.4 (Summary Evaluation) showed that Okapi BM25 and Cosine Similarity extracted different topics from the document, where the extracted topics are mentioned in the gold standard summaries. Thus, we believed that Okapi BM25 and Cosine Similarity are comparable with each other. For medium and long-length documents, Cosine Similarity showed a more accurate summary. Cosine Similarity was also used in previous work, hence we decided to use their results to compare with ours.

In the second part of the experiments, we used different settings of the related documents. The use of different values of k was discussed in the work of Wan & Xiao (2010), where they explored a range of k from 1 to 15. Similar to our results, they reported that by increasing the number of k (for $k > 10$) might not improve the summaries generated by the system. However, they did not report this with statistical significance. In Goyal et al. (2013), they did not report on the use of different numbers of expanded documents.

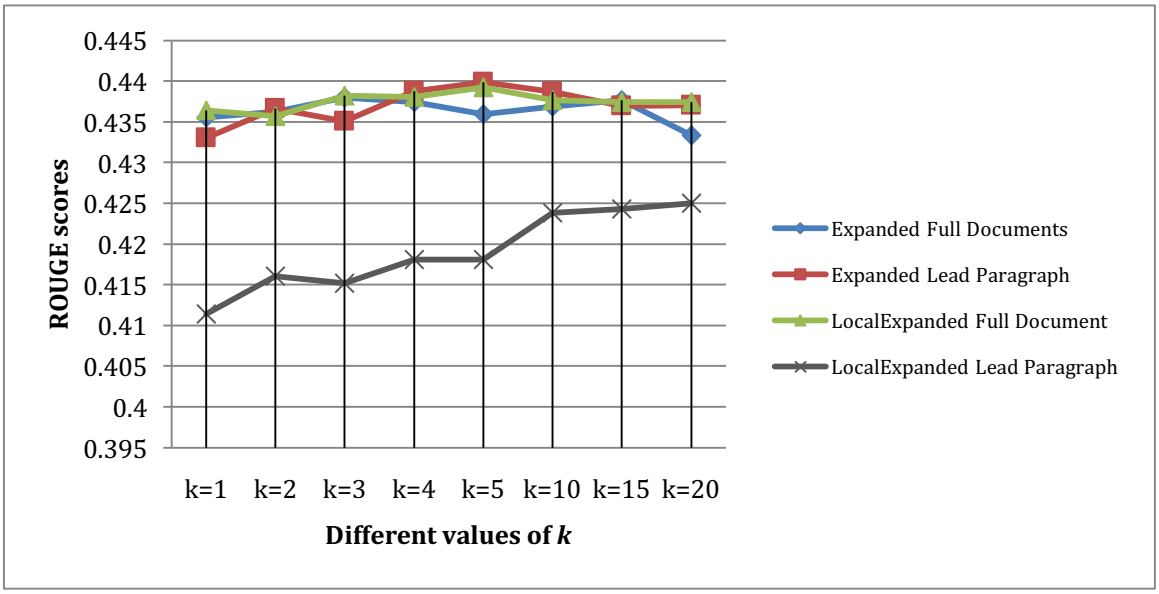


Figure 3.13 Recall ROUGE-1 scores for Expanded and Local+Expanded Document/Lead Paragraph

We also examined different lengths of documents (full document and lead paragraph) and the use of the document group dataset. Our results showed that a range of values gives a statistically similar improvement (Figure 3.13). All settings (except for the Local+Expanded Lead Paragraph) worked well for a small number of k. For the Local+Expanded Lead Paragraph, ROUGE increases as k increases. But overall, the Lead Paragraph summaries did not perform as well as we expected (except for some number of k which gives the highest ROUGE score). We assumed that more lead paragraph documents would help to create a better summary. This is also concluded by Wolf &

Gibson (2004) where their simple-paragraph algorithm (that also served as their baseline) performed poorly. For our automated summarizer, all sentences in the Local and Expanded Documents are evaluated and given a score to rank their importance. Thus, we believe that by having more sentences in a document (full document), it would improve sentence selection for a summarization tool.

It is also worth noting that when the selection of documents was restricted to those from a manually selected group, the accuracy of the Expanded Document summaries improved significantly. This was shown in the last experiments, where we used two datasets as the related document; (1) manually grouped documents only and (2) the whole dataset. However, this needs to be proved with more documents as we only had a small number of documents in the first dataset; where there are only 4 documents in each set.

We showed that Affinity Graph could improve single document summaries, similar to Wan and Xiao (2010) and Goyal et.al (2013), where the Expanded Document and Expanded+Local Document settings, gave significant improvement to the Local Document summaries. However, the improvement made by the Affinity Graph summaries was not visible in the ROUGE score as the Affinity Graph summaries did not beat the baseline summaries.

We also see that the auto-summaries generated by the Affinity Graph algorithm should be further explored based on the sentence extraction analysis discussed in Section 3.4. The diversity of the sentences extracted by the Affinity Graph algorithm was not shown in their ROUGE score. The Affinity Graph algorithm tried to identify the most discussed topics based on the relationship between the documents; and this may have contributed to different sentences from the same topic extracted as 'relevant' to the algorithm.

3.6 Conclusion and What's Next?

This study set out to reinvestigate the use of Affinity Graph for single document summaries. The discussion on the contribution includes:

1. The process of identifying a range of settings to improve single document summarization by recognising its related documents. As discussed in Section 3.3 and Section 3.4, Cosine Similarity was not improved upon by Okapi BM25 and the Language Model.
2. The use of the lead paragraph from expanded documents improved single document summarization. However, based on the ROUGE score, it does not improve significantly over the use of full document types, and our manual summary analysis also showed that the lead paragraph summaries produced summaries that were almost similar to the full document summaries. Hence, we assumed that the use of a condensed version of the document could be exploited as nowadays the information is spread widely in a short and fast way, such as in the use of Twitter. We are also interested to explore the use of other document types to support document summarization.
3. We also noticed that there is a negative correlation between document length and the ROUGE scores. The sample with longer document (see Figure 3.10) gave lower ROUGE score but (we believed) the auto-generated summaries were comparable with the human abstract summaries.

Previous work on graph-based summarization showed it is a viable approach for automatic summarization (Giannakopoulos et al., 2008; Mani & Bloedorn, 1997; Plaza et al., 2011). We believed that the Affinity Graph algorithm was able to improve the summary accuracy by including expanded documents. However, we discovered that no specific parameter is needed to determine the best setting for Affinity Graph. Each neighbourhood (Expanded, Local+Expanded, Lead Paragraph) setting has its own advantages and gave reliably good results.

We would like to explore more applications of the Affinity Graph in different domains. Hence, we identified new questions on how to build a summarizer (by applying the Affinity Graph algorithm) and make use of its expanded documents:

- Can we exploit social media to improve document summarization?
- Can the approach work with limited information (e.g. tweets)?
- How to generate summaries that take into account both information from the sentences (documents) and the interest of social users?
- Will the ROUGE scores show the same correlation pattern with document length?

Based on these questions, we developed another set of experiments to answer the second research question, and the questions mentioned above. This will be further discussed in Chapter 4.

Chapter 4
Tweet-Biased Summarization Using Affinity Graph

In the previous chapter (Chapter 3), we built on past work to explore a range of settings to improve single document summarization employing the Affinity Graph approach, using Expanded Document, and Local+Expanded Document. We examined parameter settings with different similarity measures, a number of related documents, document lengths, and the use of a manually assigned document group. We discovered that no setting or parameter was consistently better than another, based on a Recall ROUGE score. However, summaries based on expanded documents were significantly better than single document summarization.

Based on the discussion in Chapter 3, we identified our second research question (RQ2):

"Can the Affinity Graph algorithm improve single document summarization using limited length document (tweets)?"

In this chapter, we attempt to apply the Affinity Graph approach to generate tweet-biased summarization. The new Affinity Graph framework adopts the concept of generic extractive summarization for a single document. Instead of using a similar type of document with the Local

Document, we used Tweets (Figure 4.1). We believe that the tweets would be a good representation of condensed and limited information to support the local document.

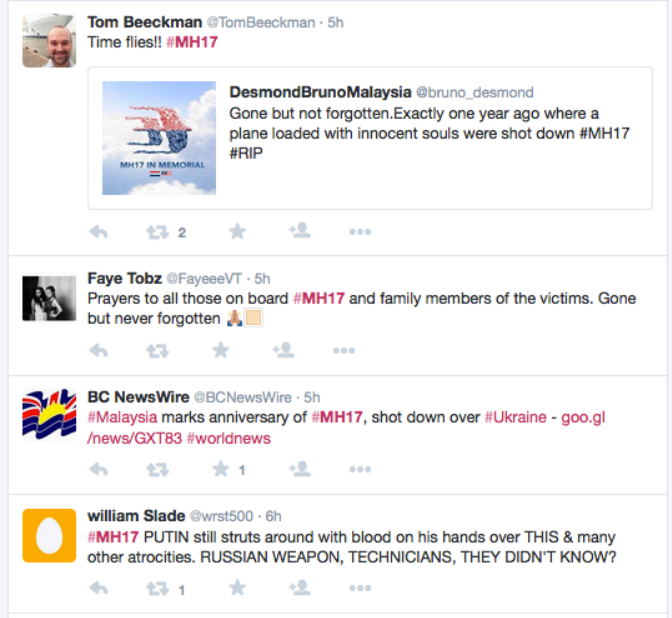


Figure 4.1 Example of tweets

The chapter contributes:

1. A new Affinity Graph framework that includes social media content (so-called Expanded Tweets). We tested different Affinity Graph settings. However, unlike the work described in Chapter 3, we found that there are only small differences between the parameters tested.
2. A new dataset is introduced to test the tweet-biased Affinity Graph framework.
3. A system to build the dataset (Sentence Extraction System –SESys) is described.
4. Further analysis on variations of ROUGE score and document length.

4.1 Background Work

Previous researchers have studied social media summarization using different approaches: graph algorithms (M. Hu et al., 2007; Yang et al., 2011), topic modeling (Gao et al., 2012), and novelty detection algorithms (Parapar et al., 2010; Yulianti, 2013). Boydell & Smyth (2007), P. Hu, Ji, et al.,

(2011a) and Park et al. (2008) applied their summarization techniques to social bookmarking websites to produce a higher quality document summaries, when compared with the baseline systems/other benchmark summarization tools and manual summaries by human evaluators.

Twitter (in particular) has gained much attention (Gao et al., 2012; Kothari et al., 2013; Nichols et al., 2012; Ritter et al., 2010; Sharifi et al., 2010). Recent work looking at users of social media, such as Twitter, suggests that users often comment on parts of Local (web) Documents that are considered important or interesting. We assumed that this information could be used to select important sentences from a web document and thus, would improve any summary of it.

4.1.1 Affinity Graph for Tweet-Biased Summarization

Our framework was developed using an Affinity Graph approach, as in previous experiments, where we measured the similarity between each related tweet and Local (web) Documents. For this experiment, our main tasks are:

1. **To build an Affinity Graph of a local document and its related tweets.**

   We identified the related tweets of a set of local documents. Based on results in Chapter 3, we applied the Cosine similarity technique to calculate pairwise relationships. We used the Lemur Toolkit to calculate the Cosine value.

2. **Summarizing the local document using different settings of the related tweets.**

   We applied the Affinity Graph algorithm to generate summaries for the Local Document. We examined how informative the related tweets are to improve summaries. Here we are also interested to see if the combined information from the sentences (documents) and interest of the social users (tweets) would be able to improve the document summaries.

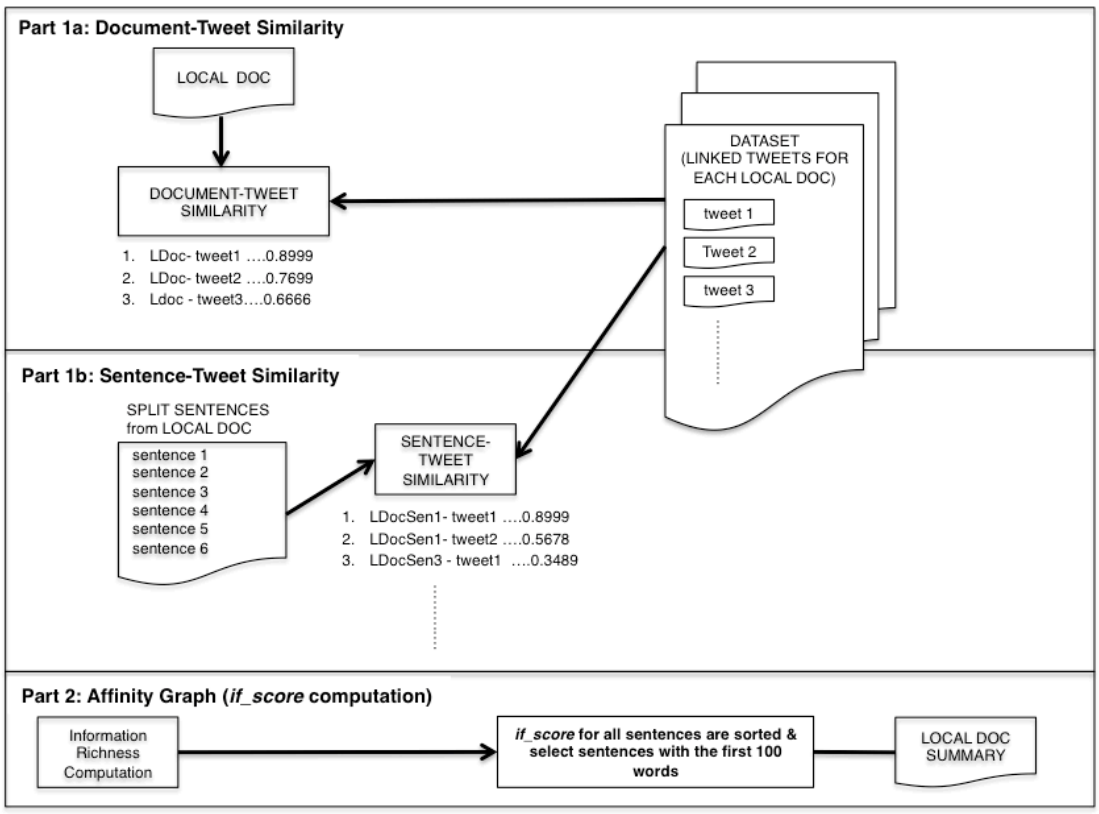


Figure 4.2: Affinity Graph framework for Tweet-Web dataset

Figure 4.2 is based on the Affinity Graph approach discussed in Chapter 3, but we changed our neighbourhood setting, as follows:

1. **Local Document** use only information from within the local document itself (Informativeness from Content).
2. **Expanded Tweets** use information from the related tweet of the local documents (Social Content).

4.2 Experiment Setup

For these experiments, we aimed to explore the use of social information as additional clues to extract sentences from web documents. Our framework was developed using an Affinity Graph approach, as in previous experiments, where we measured the similarity between each related tweet and the web

documents. In this section, we will focus the discussion on the development of the new dataset and the new framework.

4.2.1 Tweet-WebDoc Dataset

For this experiment, we applied our summarization system to a new dataset developed by Yulianti (2013). The Tweet-WebDoc dataset was based on the TREC 2011 Microblog track, which held 16 million tweets, collected from January 23rd to February 8th, 2011. The pre-processing of the tweets are as in Figure 4.3. Yulianti (2013) extracted 15,167,481 tweets (with textual information) from the TREC2011. At the end of the pre-processing of the dataset, were left 493 web (or local) documents with related Expanded Tweets (minimum 10 tweets) that hold links to each of the documents.

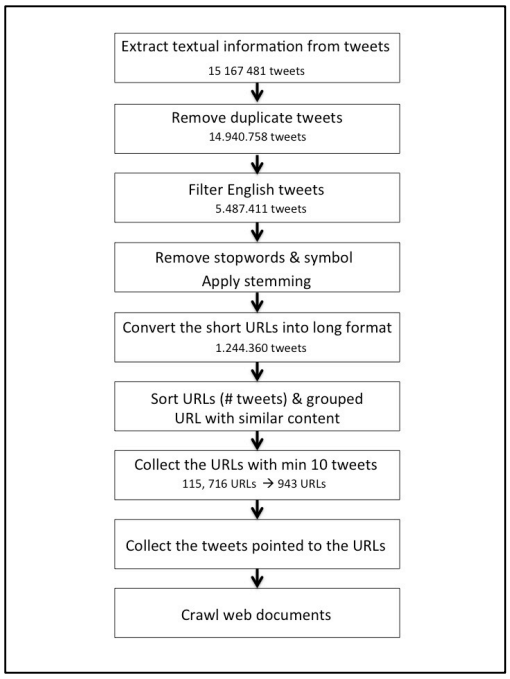


Figure 4.3: Pre-processing of Tweet-WebDoc dataset

Table 4.1 shows URL categories of the web documents (using the FortiGuard[1] web filtering tool). FortiGuard's Web Filtering categorized the websites into six main groups, and each of the

[1] http://www.fortiguard.com/ip_rep.php

websites was assigned based on their dominant Web content (FORTINET, 2007, Yulianti, 2013). Most URLs were from News and Media.

Table 4.1: URL category

Category	Num Of Doc
News and Media	349
Information Technology	87
Personal Websites and Blogs	18
Reference	9
Sports	8
Business	7
Entertainment	4
Finance and Banking	2
General Organizations	2
Political Organizations	2
Education	1
Health and Wellness	1
Newsgroups and Message Boards	1
Peer-to-peer File Sharing	1
Streaming Media and Download	1

Table 4.2 shows the Top 10 URL domain, with the most number of tweets linked to its article. Here, we can see that most domains are news websites: Mashable, CNN, BBC, Huffington Post, Guardian, New York Times, and Al-Jazeera. These domains also have a high number of tweets pointing to one of its articles. Note, Al-Jazeera is in the Top10 domain list, most likely because the Egypt revolution was dominating the news at the time the tweets were collected.

Table 4.2: Top 10 Domain

Domain	Num of URL	The most number of tweets
mashable.com	196	43
techcrunch.com	50	27
cnn.com	37	36
bbc.co.uk	17	39
huffingtonpost.com	11	22
aljazeera.com	11	28
guardian.co.uk	10	38
nytimes.com	8	24
helium.com	8	15
wsj.com	7	16

Figure 4.4 shows the distribution of the number of tweets pointing to its Local Document. Here we can see that the most Local Documents have a minimum of 10 tweets.

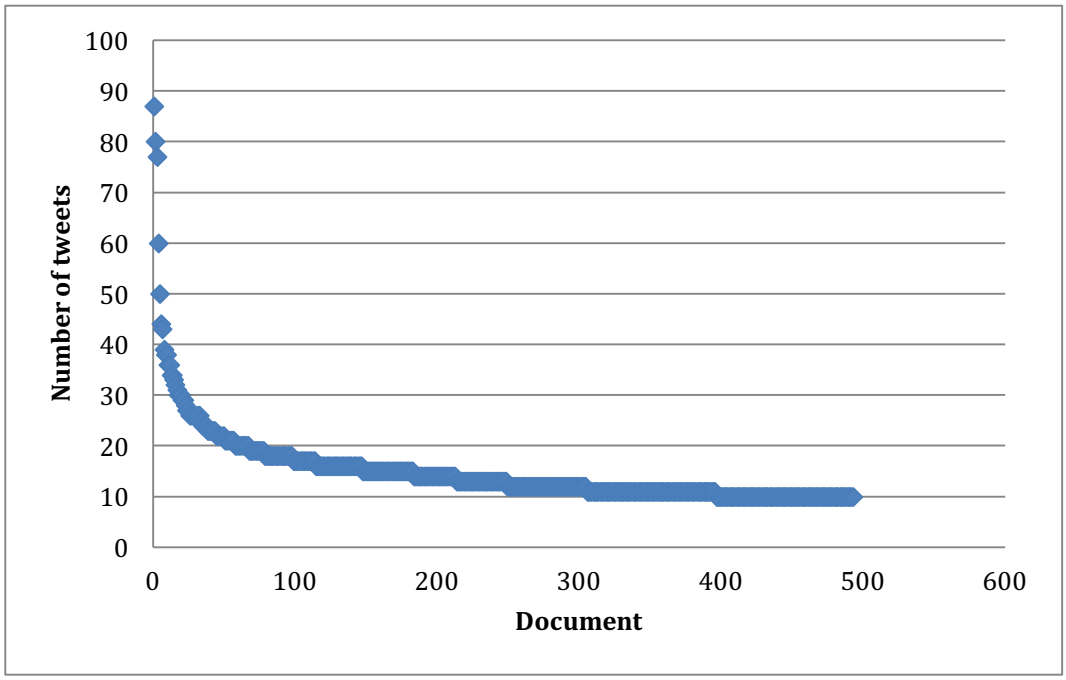


Figure 4.4: The number of tweets for each documents in the dataset (Yulianti, 2013; Yulianti et al., 2015)

4.2.2 Affinity Graph Setup

Since tweets are already related to the Local Documents (through links), we ran the affinity graph algorithm in the same setting as the manual group document described in Chapter 3. We defined Document/Tweet similarity in two ways. First assuming the tweets are part of one document, thus the value of λ is set to 1 (Figure 4.5). Second assuming each tweet is a document on its own. Thus, for each of the document – tweets are defined as:

$$\lambda = sim_{doc-tw}(d_i, t_j)) \quad (4.1)$$

Equation 4.1 is set to the affinity value calculated between the Local Document and its related tweets (see Figure 4.6). Here each of the tweets will have its own affinity value (λ) and this will be used to create matrix M.

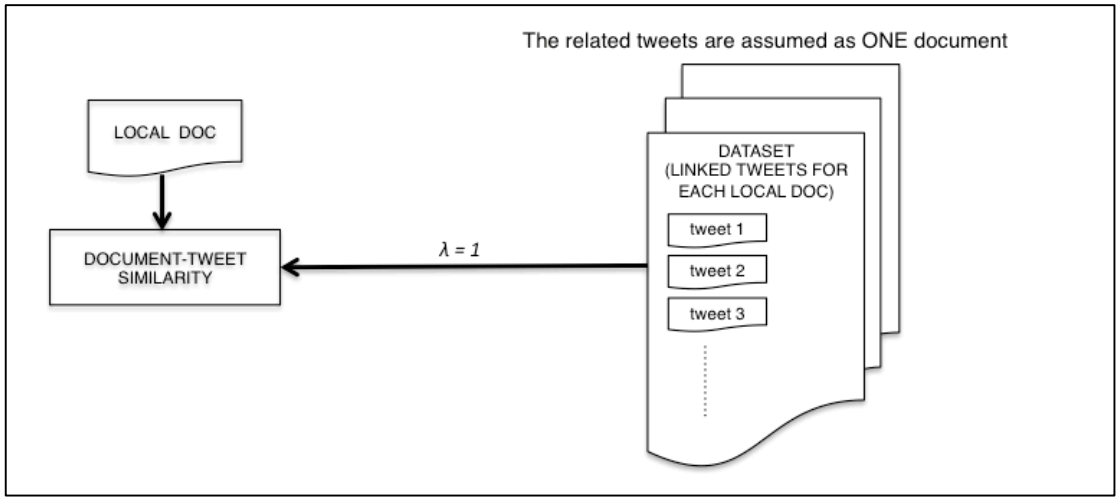


Figure 4.5 Document – Tweet relationship as one document.

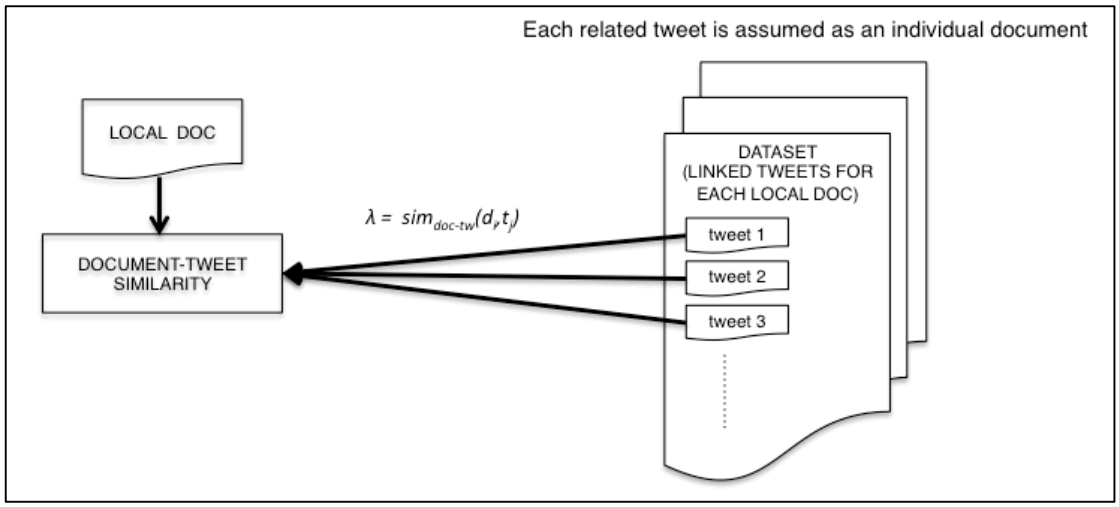


Figure 4.6 Document – Tweet relationship where each tweet is viewed as a separate document.

Once we have defined the affinity score (λ), the matrix M is created as in Equation 4.2:

$$M_{i,j} = \begin{cases} \lambda \times sim_{sen-tw}(s_i, s_j), & if\ i \neq j \\ 0 & otherwise \end{cases} \quad (4.2)$$

This is different from Equation 3.1, where the sentences from the Local Document are defined by (s_i) and a related tweet is defined by (t_j). We normalized M with the same Equation 3.2 and calculated the informativeness score (*if_score*) as in Equation 3.3. The *if_score* for each sentence in Local Document was sorted and extracted until the summary word limit was reached.

We identified the following neighbourhood environments:

1. Expanded Tweets using Document and Sentence Similarity (EXP)
2. Expanded Tweets using Sentence Similarity, where $\lambda = 1$ (EXP_S)[2]
3. Top 10 Expanded Tweets (T10.EXP)
4. Local Document and Expanded Tweets (LD.EXP).

[2] EXP_S, *S* is the abbreviation for Sentence similarity because in this relationship, only sentence – tweets relationship gives the similarity values.

The first two settings are used to examine if Document – Tweet similarity is affected by the document and tweet relationship. For the first setting (EXP), the Document - Tweet similarity has the same query files as in Figure 3.7 and because we used LEMUR to calculate the Cosine Similarity, we need to build the index for the tweets as the datafiles. This is the main difference for the similarity process in this framework, where only the tweets will be used to index and then calculate the similarity value to the documents.

Sentence-Tweet similarity's query file is as in Figure 3.8 and this is applied to all the sentences – tweet relationships. As mentioned in section 4.2, we were interested to see if using the minimum number of tweets would able to improve the summary. Thus, we selected the top 10 tweets (based on the Document-Tweets similarity values) and used the tweets to generate a summary (T10.EXP). We also used information from the Local Document only as in the Local Document setting in Chapter 3. For this, we defined the setting as LOCAL. In another expanded setting, we merged the Local document with its tweets (LD.EXP), to see if by combining both pieces of information we could improve the summary compared to other settings.

We also created two baseline systems. The first (Baseline 1) consists of the first 100 words of a document, the same definition as the baseline used in Chapter 3. The second (Baseline 2), contains the first 100 words from the first sentence of each paragraph of a document.

We also compared to the Tweet-based summarization system (TBS) (Yulianti et al., 2015) which used a different summarization approach proposed by Parapar et al. (2010). In Yulianti et al's work, they developed two summarization systems; a Generic Summarization (GS_{sn}) that used only information from the Local Document, and TBS, which used the related tweets to generate a summary for the Local Document[3].

TBS is based on the ranking of related tweets and the sentences from the document to be summarized. Tweets were ranked based on their relevance to the Local Document. Then a novelty

[3] GS_{sn} uses the same information as our LOCAL summarization setting and TBS uses the same neighbourhood as our EXP.

detector system removed redundant tweets. Yulianti (2013) and Yulianti et al. (2015) selected and defined the top 30% of the ranked tweets to be 'novel tweets'. These tweets were then combined to form a new query. The process was repeated for sentences from the Local Document. This time, the sentences were ranked based on the 'new query' and the novelty detector system was reapplied[4].

4.2.3 ROUGE Evaluation

We used ROUGE (with the same parameter setting in Chapter 3) to evaluate the summary. Thus, we need to create gold standard summaries (human-generated summaries) for the Tweet-WebDoc dataset. Creating such summaries for evaluation is commonly practised (Inouye & Kalita, 2011; Liu & Liu, 2010).

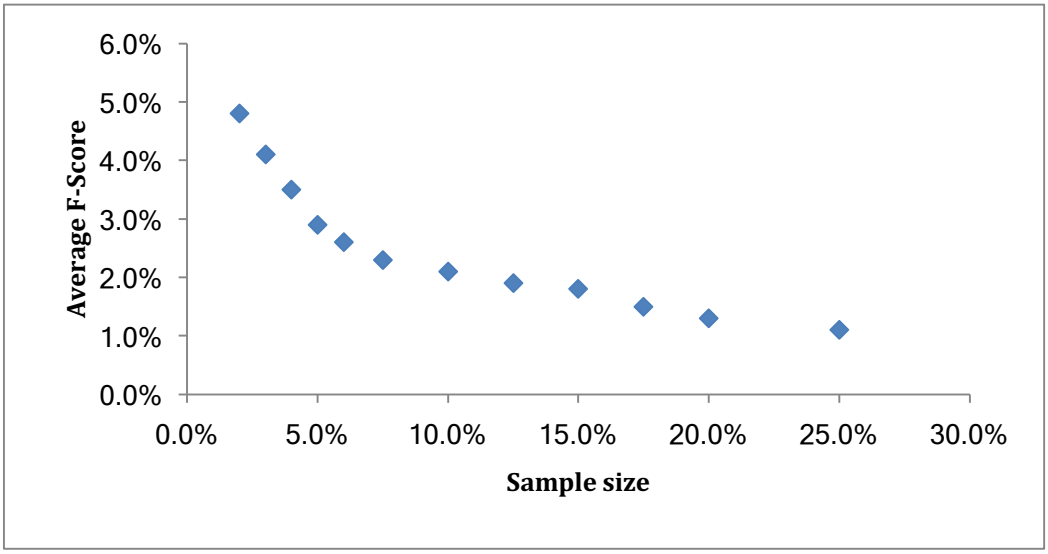


Figure 4.7: The Difference between the Average and Overall F-score

We analysed DUC2002 ROUGE scores from Chapter 3 to estimate the number of summaries needed for evaluation. Figure 4.7 shows the relationship between the average F-score for the sample

[4] The similarity between TBS and our Affinity Graph algorithm is on the use of tweets as the related documents, however the implementation is different. We define the relationship between the tweets and document/sentences through Cosine Similarity. We then integrate the tweet-document and tweet-sentence relations into an affinity algorithm (Equation 4.1) to obtain an affinity score (λ). The ranking of the extracted sentences is based on λ. The Affinity Graph smoothes the relationships between the sentences, as it used (1) the relationship between the sentences and the tweets (from cosine similarity), and (2) the informativeness of the sentence (from the affinity scores).

and the average F-score for the whole dataset. Here, we repeatedly sampled the F-scores of different sample sizes. We measured the difference between an average sample F-score and the overall average of the F-score. It is assumed that if we took 10% sample from the dataset, we would get 2% average variance F-score for the whole dataset. Based from this, we estimated that using 10% sample to further analysed the summary evaluation would be good enough to represent the new dataset. Thus, we randomly chose 55 documents[5] (~ 10% of 493 documents) to be manually summarised for the dataset[6]. The category of the 55 documents is shown in Table 4.3.

Table 4.3: Categories for 55 documents

Category	**Number documents**
News and Media	41
Information Technology	7
Personal Websites and Blogs	3
Business	1
Reference	1
Sports	1
Streaming Media and Download	1
Total	55

Considering document length, the longest sampled document has 694 sentences, the shortest, 4. For related tweets, the sample dataset has a range of 10-80 tweets. Based on this, we believe that the 55 documents would represent the whole dataset, and we would be able to evaluate the Affinity Graph algorithm for tweet-biased summarization for the single document.

[5] To randomize the documents, we used a Randomizer tool (https://www.randomizer.org/)

[6] Initially we choose 5% of the documents from the dataset (30 documents), however we added 25 more as suggested in the Power Analysis. This will be explained in section 4.3.1.

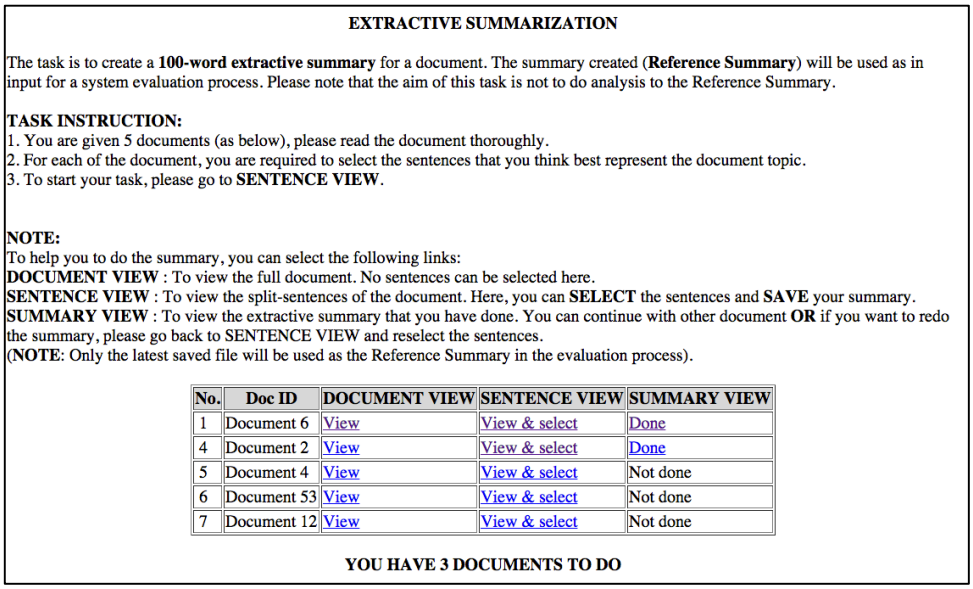
4.2.4 Sentence Extraction System for Reference Summaries (SESys)

For the reference summaries in DUC2002 (Chapter 3), NIST provided two summaries for each document, manually written as abstracts of the document. Our main focus is the sentence extraction. Thus, rather than asking a human summarizer to abstract a document, the task was to identify and select sentences from the local document that would contribute to a summary.

We asked 22 postgraduate students from universities around Melbourne, Victoria (e.g. RMIT, Melbourne University, Victoria University and Swinburne University) to select sentences from documents that they think are important and relevant to those document. The method of having non-expert volunteers as participants to generate reference summaries is also discussed by Gao et al. (2012), M. Hu et al. (2008), Inouye & Kalita (2011) and Liu & Liu (2010).

We contacted the students via email, instructing them to read the documents and select sentences that best represent the documents (Figure 4.8). Each participant was given five documents, and each document was summarised by two participants. We asked the work to be done in 1-2 days, but some responses were received two to three weeks after the invitation emails. In total, 110 summaries were created.

Figure 4.8 shows the main page for a system (SESys) we developed to gather selected sentences. Here, the description and instructions for the task are presented. The lists of the documents for the participants can also be viewed on the main page. The description of the features of the system is also displayed to provide an overview of the tasks for the participants.



EXTRACTIVE SUMMARIZATION

The task is to create a **100-word extractive summary** for a document. The summary created (**Reference Summary**) will be used as in input for a system evaluation process. Please note that the aim of this task is not to do analysis to the Reference Summary.

TASK INSTRUCTION:
1. You are given 5 documents (as below), please read the document thoroughly.
2. For each of the document, you are required to select the sentences that you think best represent the document topic.
3. To start your task, please go to **SENTENCE VIEW**.

NOTE:
To help you to do the summary, you can select the following links:
DOCUMENT VIEW : To view the full document. No sentences can be selected here.
SENTENCE VIEW : To view the split-sentences of the document. Here, you can **SELECT** the sentences and **SAVE** your summary.
SUMMARY VIEW : To view the extractive summary that you have done. You can continue with other document **OR** if you want to redo the summary, please go back to SENTENCE VIEW and reselect the sentences.
(**NOTE**: Only the latest saved file will be used as the Reference Summary in the evaluation process).

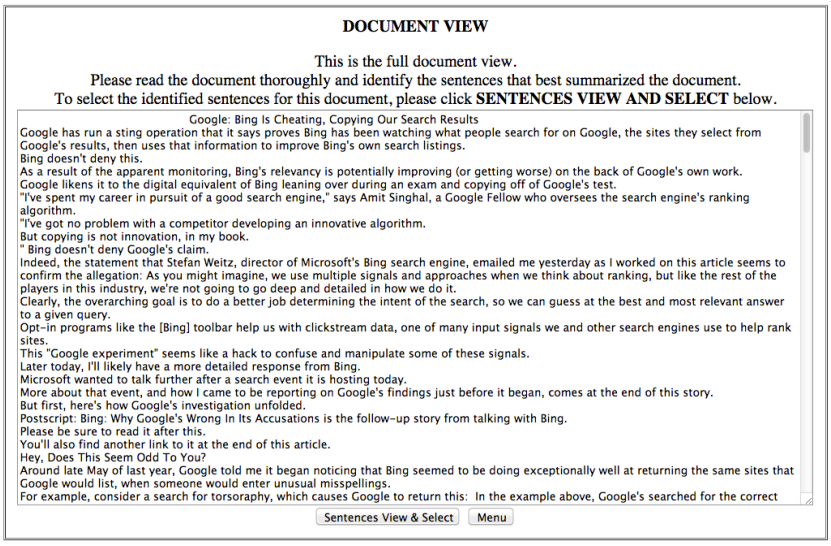
No.	Doc ID	DOCUMENT VIEW	SENTENCE VIEW	SUMMARY VIEW
1	Document 6	View	View & select	Done
4	Document 2	View	View & select	Done
5	Document 4	View	View & select	Not done
6	Document 53	View	View & select	Not done
7	Document 12	View	View & select	Not done

YOU HAVE 3 DOCUMENTS TO DO

Figure 4.8: Screen Shot of the SESys (Main Menu).

The system has three main features:

1. DOCUMENT VIEW: participants can view the full documents. No sentences can be selected from here (Figure 4.9). The Document View is opened in a separate window so that the participants can directly read the full document and select the relevant sentences (in another window – Figure 4.10). We believed that this is useful, especially for longer documents.



DOCUMENT VIEW

This is the full document view.
Please read the document thoroughly and identify the sentences that best summarized the document.
To select the identified sentences for this document, please click **SENTENCES VIEW AND SELECT** below.

Google: Bing Is Cheating, Copying Our Search Results
Google has run a sting operation that it says proves Bing has been watching what people search for on Google, the sites they select from Google's results, then uses that information to improve Bing's own search listings.
Bing doesn't deny this.
As a result of the apparent monitoring, Bing's relevancy is potentially improving (or getting worse) on the back of Google's own work.
Google likens it to the digital equivalent of Bing leaning over during an exam and copying off of Google's test.
"I've spent my career in pursuit of a good search engine," says Amit Singhal, a Google Fellow who oversees the search engine's ranking algorithm.
"I've got no problem with a competitor developing an innovative algorithm.
But copying is not innovation, in my book.
" Bing doesn't deny Google's claim.
Indeed, the statement that Stefan Weitz, director of Microsoft's Bing search engine, emailed me yesterday as I worked on this article seems to confirm the allegation: As you might imagine, we use multiple signals and approaches when we think about ranking, but like the rest of the players in this industry, we're not going to go deep and detailed in how we do it.
Clearly, the overarching goal is to do a better job determining the intent of the search, so we can guess at the best and most relevant answer to a given query.
Opt-in programs like the [Bing] toolbar help us with clickstream data, one of many input signals we and other search engines use to help rank sites.
This "Google experiment" seems like a hack to confuse and manipulate some of these signals.
Later today, I'll likely have a more detailed response from Bing.
Microsoft wanted to talk further after a search event it is hosting today.
More about that event, and how I came to be reporting on Google's findings just before it began, comes at the end of this story.
But first, here's how Google's investigation unfolded.
Postscript: Bing: Why Google's Wrong In Its Accusations is the follow-up story from talking with Bing.
Please be sure to read it after this.
You'll also find another link to it at the end of this article.
Hey, Does This Seem Odd To You?
Around late May of last year, Google told me it began noticing that Bing seemed to be doing exceptionally well at returning the same sites that Google would list, when someone would enter unusual misspellings.
For example, consider a search for torsoraphy, which causes Google to return this: In the example above, Google's searched for the correct

Sentences View & Select | Menu

Figure 4.9: Screen Shot of the DOCUMENT VIEW in SESys.

2. SENTENCE VIEW: participants can read and select sentences (Figure 4.10). The system will auto-calculate the number of words in the selected sentences. Once the 100-word limit is reached, participants can save their selection. If they selected more than the word limit, they are allowed to save, but the system will mention that only the first 100 words of their selection will be used as the reference summary.

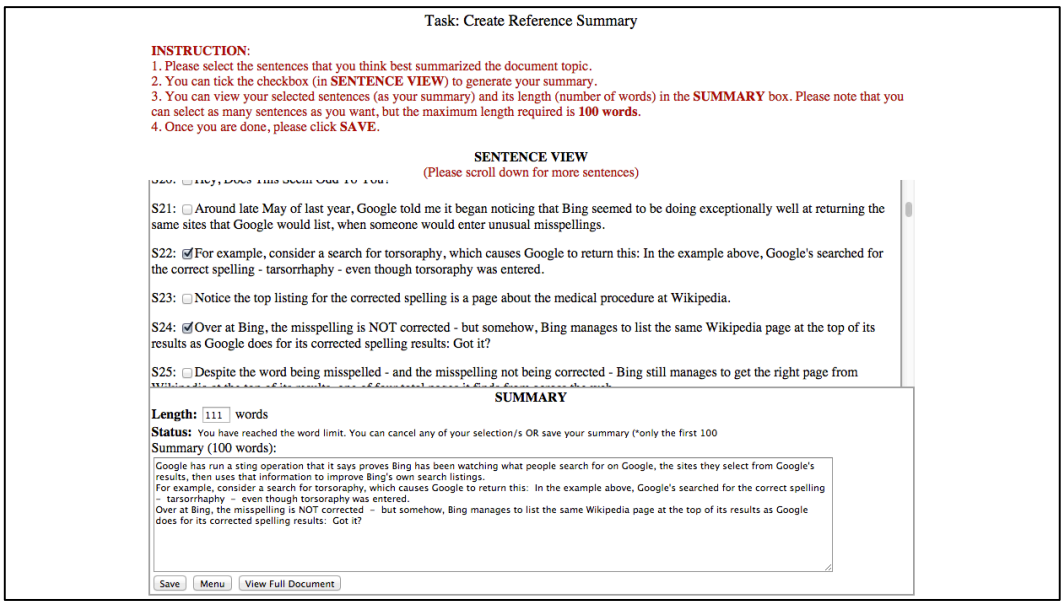


Figure 4.10: Screen Shot of the SENTENCE VIEW in SESys.

3. SUMMARY VIEW: participants can review their summary (Figure 4.11). Participants are allowed to change their sentence selection by going back to the SENTENCE VIEW. Only the last saved summary will be used as the reference summary.

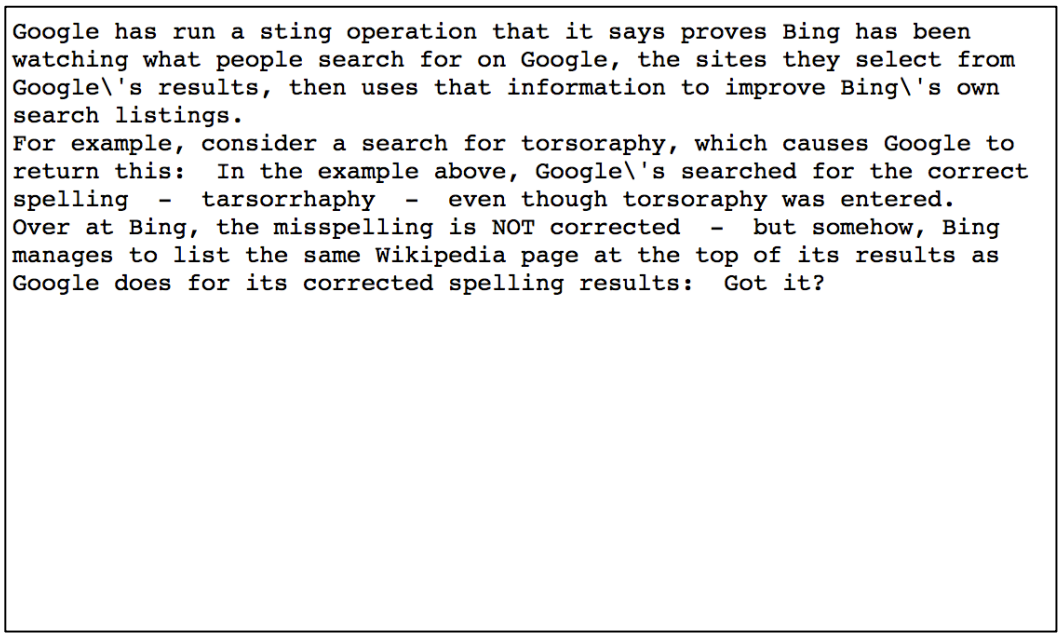


Figure 4.11: Screen Shot of the SUMMARY VIEW in SESys.

4.2.5 Kappa Agreement for Human Summarizers

We measured the human summariser (i.e. sentence selection) agreement using the average term-level Kappa ratio (κ) for all reference summaries. The κ represents the percentage of agreement between the raters. The maximum agreement is 1, but a perfect agreement is rare. Table 4.4 shows one possible interpretation of κ (Viera & Garret, 2005).

Table 4.4 Kappa (κ) agreement interpretation

Kappa (κ)	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

For our reference summaries, we calculated a κ agreement of 0.33 (Yulianti et al., 2015). Based on Table 4.4, this shows that our human summarizer gave a fair agreement in extracting the sentences for our reference summary. This result is comparable with previous studies on document summarization (Hirohata, Shinnaka, Iwano, & Furui, 2005; Keikha, Park, & Croft, 2014).

4.3 Results

We explored the accuracy of the different summarisers. We started by examining different affinity graph settings.

In this experiment (shown in Table 4.5), only Baseline 1 shows significantly different scores when compared with all other settings (for all ROUGE-1 and ROUGE-2 Recall scores). For Baseline 2, it was significantly better on two measures (Precision ROUGE-1 and ROUGE-2) of EXP_s.

Table 4.5: ROUGE scores for Baseline 1, Baseline 2 and the Tweet-biased summaries

	ROUGE-1		ROUGE-2	
	RECALL	PRECISION	RECALL	PRECISION
BASELINE 1	0.616	0.480	0.466	0.360
BASELINE 2	0.519	0.460	0.357	0.328
LOCAL	0.523*	0.439*	0.328*	0.276*#
EXP_s	0.499*	0.418*#+	0.307*	0.258*#+
EXP	0.528*	0.452	0.353*	0.305*
T10. EXP	0.523*	0.445*	0.342*	0.294*
LD.EXP	0.526*	0.430*	0.333*	0.269*#
GS_{sn}	0.478*+~	0.374*+~	0.269*+	0.205*#+~
TBS	0.555*	0.431*	0.377*	0.290*

** statistically significant (p-value < 0.05) compared to BASELINE 1 summaries*
\# statistically significant (p-value < 0.05) compared to BASELINE 2 summaries
\+ statistically significant (p-value < 0.05) compared to EXP summaries
~ statistically significant (p-value < 0.05) compared to LOCAL summaries

For the four document-tweets relationship settings: EXP, EXP_s, T10.EXP, Local, and LD.EXP. Only small differences in ROUGE scores were found. We performed paired t-tests and found that, with one exception, none of the significance tests showed that one setting is better than the

other. The exception was EXP when compared to EXP_S in its Precision score for ROUGE-1 and ROUGE-2.

This shows that different Affinity Graph settings had no significant effect on the generated summaries. Thus, based on the ROUGE scores, the generated summaries are similar to each other. Table 4.5 also shows the comparison with TBS and GS_{sn}. We found the ROUGE scores for GS_{sn} were significantly worse[7], however, no significant difference was found between TBS and the other summarizers.

4.3.1 Power Analysis

As discussed in section 4.3.1 and section 4.3.2, most of the comparisons were not significant. Therefore, we conducted a post-hoc power analysis to the Recall ROUGE-1 scores for all settings.

Power analysis determines the ability to find a difference/effect of a study given that the difference/effect really exists (Webber, Moffat, & Zobel, 2008). There are two types of errors that are likely to be observed in a statistical power analysis. Type I error (false positives) is the chance that one has incorrectly rejected a null hypothesis – detecting an effect when it actually does not happen. While a Type II error (false negative) fails to detect an effect that actually happens. A higher power would mean that there is a greater chance to find statistical significance when it happens and thus able to avoid a Type II error.

The sample size, the alpha level, and the effect size will determine if the study is 'powerful' enough to produce a statistically significant difference. A sample size (symbolised by *n*) is the number of data (usually randomly selected from the pool of dataset) that we used to test our hypothesis. A larger sample of data would generally give us a higher power, and would be easier for us to reject the null hypothesis. Whereas the Alpha level (α) is the probability of a Type 1 error (or the error rate) and is usually set to 0.05 or 0.1.

[7] We applied the paired t-test to the ROUGE results.

An Effect Size (ES) is a measure to indicate an impact of the factors that affects the outcome of the study. The ES can be measured by calculating the mean difference between the two variables (*d*), the correlation between the variables or the regression coefficient (*r*) (Cohen, 1988). Cohen (1988) has defined that an ES of 0.2 represents a small effect, 0.5 represents medium effect, and 0.8 shows a large effect. A Small ES would likely happen due to uncontrollable variables that make the effect hard to be detected. A Medium ES would show that the effect can be visible for certain cases, but for some, the effect would still be considered as small. However, for a Large ES, the separation of effects between different results is easily visible and accepted.

There are two main types of power analysis; a priori and a post-hoc analysis. A priori power analysis is usually conducted before the data collection stage to estimate a sample size for the study. For post-hoc power analysis, it is done after a study has been completed. The post-hoc power analysis is based on the sample size and effect size to determine the power of the study (Faul, Erdfelder, Lang, & Buchner, 2007).

As discussed in Section 4.2.3, we calculated our sample size based on the results derived from our DUC2002 experiments. Figure 4.7 suggested that a 5% sample (~30 documents) would be enough to show a significant difference in ROUGE. However, we decided to use 55 documents (10%) for our experiment. The power analysis shows that the experiment was likely to have a Type II (false negative) error. We calculated the power values for all of the settings using a sample size n=55 and error rate α =0.05. We also calculated the Effect Size (*d*) (Cohen, 1988) based on the paired t-test:

$$d = \frac{|m_A - m_B|}{\sigma} \qquad (4.2)$$

We defined $m_A - m_B$ (in Equation 4.2) as the difference in means between two paired settings and σ as the standard deviation of one of the settings (we assumed that the two settings were equal as described by Cohen (1988)). Table 4.6 shows the *d* and power for all the paired settings based on the Recall ROUGE-1 results as in Table 4.5.

Based on the power analysis results (Table 4.6), we determined that the power for our tweet-biased summarization system that is far less than the recommended statistical power of 0.8. This result shows a high probability of a Type II error. However, for the paired settings of GS with EXP and LOCAL shows a higher statistical power (0.6 and 0.5 respectively) because the paired settings gave a significant difference as discussed in Section 4.3.2.

Table 4.6 The Effect Size (d) and Power for all paired settings

	Effect Size (d)	**Power**
LOCAL and EXP	0.02	0.1
LOCAL and EXP_s	0.11	0.1
LOCAL and T10.EXP	0.05	0.1
LOCAL and LD.EXP	0.06	0.1
EXP and EXP_s	0.17	0.2
EXP and T10.EXP	0.03	0.1
EXP and LD.EXP	0.01	0.1
EXP and T10.EXP	0.14	0.2
EXPs and LD.EXP	0.17	0.2
T10.EXP and LD.EXP	0.02	0.1
GS and EXP*	0.30	0.6
TBS and EXP	0.15	0.2
LOCAL and TBS	0.11	0.1
LOCAL and GS*	0.27	0.5

**the settings show a significant difference in the paired t-test*

We believed that the low statistical power (0.1 and 0.2) was caused by the small sample. We also applied a post-hoc power analysis, where it suggested to increase *n* to 2000 in order to achieve a large power of 0.8. However, we decided not to increase the number of sample for the experiment as this problem has been discussed by Goodman & Berlin (1994) and Trout, Kaufmann & Kallmes (2007). They stated that post-hoc power calculation to determine the 'right' sample size would not help to achieve better power.

4.4 Summary Evaluation

In Chapter 3 (Section 3.4), we see that different summarization systems generated different summary content. Consequently, we further analysed the summaries by: (1) manually looking at the extracted sentences in the summaries, (2) analysing ROUGE score correlations, and (3) ROUGE score variations.

4.4.1 Manually Examining Summaries

In Figures 4.12 - 4.15, we show baseline and reference (manual) summaries as well as examples of their related tweets from four different documents drawn from the collection:

- DocID 2 was identified as the document with a high number of sentences (208 sentences) and a high number of related tweets (80 tweets). DocID 2 is a blog article.
- DocID 311 has a high number of sentences (694 sentences) and a low number of tweets (11 tweets). It is from an online magazine.
- DocID 55 has a low number of sentences (6 sentences) and a high number of tweets (21 tweets).
- DocID 426 has a low number for both sentences in the document (9 sentences) and tweets (10 tweets).

In Figure 4.12, both reference summaries have the same topic (*Google caught Bing copying their results*). In the summaries generated by the Affinity Graph, only those generated using the local document (LOCAL and LD.EXP) gave a high Recall ROUGE score (0.65).

<table>
<tr><th colspan="2">DOCUMENT (DocID 2)
TITLE: Google: Bing Is Cheating, Copying Our Search Results</th></tr>
<tr><th colspan="2">BASELINE 1
Recall ROUGE-1: 0.667</th></tr>
<tr><td colspan="2">Google has run a sting operation that it says proves Bing has been watching what people search for on Google, the sites they select from Google's results, then uses that information to improve Bing's own search listings. Bing doesn't deny this. As a result of the apparent monitoring, Bing's relevancy is potentially improving (or getting worse) on the back of Google's own work. Google likens it to the digital equivalent of Bing leaning over during an exam and copying off of Google's test. "I've spent my career in pursuit of a good search engine," says Amit Singhal, a Google Fellow….</td></tr>
<tr><th colspan="2">BASELINE 2
Recall ROUGE-1: 0.660</th></tr>
<tr><td colspan="2">Google has run a sting operation that it says proves Bing has been watching what people search for on Google, the sites they select from Google's results, then uses that information to improve Bing's own search listings. As a result of the apparent monitoring, Bing's relevancy is potentially improving (or getting worse) on the back of Google's own work. "I've spent my career in pursuit of a good search engine," says Amit Singhal, a Google Fellow who oversees the search engine's ranking algorithm. Bing doesn't deny Google's claim. As you might imagine, we use multiple signals and approaches when we …</td></tr>
<tr><th colspan="2">TWEET EXAMPLES</th></tr>
<tr><td colspan="2">Google: Bing Is Cheating, Copying Our Search Results: Google has run a sting operation that it says proves Bing ...
RT @dannysullivan : Google: Bing Is Cheating, Copying Our Search Results
Google: Bing Is Cheating, Copying Our Search Results Great article by @dannysullivan
Here's something interesting >> Google: Bing Is Cheating, Copying Our Search Results
RT @sengineland : Google: Bing Is Cheating, Copying Our Search Results
Google: Bing Is Cheating, Copying Our Search Results: Comments
Bing might be benefiting from google search results for longtail keywords;
Are you a BING fanboy? Well, got some bad news for you...
"Google: Bing Is Cheating, Copying Our Search Results"- Search engines spying on you to improve their SERPs.
If you can't innovate, duplicate, right?
BING now stands for "Bing Is Now Google";... via @sengineland
Microsoft's Bing copies google search results, uses Internet Explorer to track user's data please RT
#Microsoft are dirty lying cheaters. If you can't 'em, steal their stuff? Makes their ads seem kind of ridiculous...
RT @mattcutts : BREAKING: Bing cheating, copying Google's results? You have to read this: RT GENUIS
@stephanierice : Bing cheats by using Google image searches - here is the proof
Use Bing, get the EXACT same results as Google gives! Now THERE'S a unique value prop to be proud of!
Google ran an amazing sting operation to prove that Bing is copying its results.</td></tr>
<tr><th colspan="2">REFERENCE (MANUAL) SUMMARIES</th></tr>
<tr><th>Summary 1</th><th>Summary 2</th></tr>
<tr><td>[1] Google has run a sting operation that it says proves Bing has been watching what people search for on Google, the sites they select from Google's results, then uses that information to improve Bing's own search listings.
[2] Bing doesn't deny this</td><td>[1] Google has run a sting operation that it says proves Bing has been watching what people search for on Google, the sites they select from Google's results, then uses that information to improve Bing's own search listings.
[2] Around late May of last year, Google told me it began noticing that Bing seemed to be doing exceptionally well at returning the same sites that Google would list, when someone would enter unusual misspellings.
[3] Despite the word being misspelled - and the misspelling not being corrected - Bing still manages to get the right page from Wikipedia at the top of …….</td></tr>
<tr><td colspan="2"></td></tr>
<tr><th colspan="2">LOCAL
Recall ROUGE-1: 0.646</th></tr>
<tr><td colspan="2">[1] Suggested Sites is one of likely ways that Bing may have been gathering information about what's happening on Google.
[2] Google has run a sting operation that it says proves Bing has been watching what people search for on Google the sites they select from Google's results then uses that information to improve Bing's own search listings.
[3] These are just some of the signals that both Bing and Google use.
[4] By no means did Bing have exactly the same search results as Google.
[5] It strongly suggests that Bing was copying Google's results by watching what some people do at Google via Internet Explorer ...</td></tr>
<tr><th>EXP_S
Recall ROUGE-1: 0.403</th><th>LD.EXP
Recall ROUGE-1: 0.646</th></tr>
<tr><td>[1] If its not illegal is what Bing may be doing unfair somehow cheating at the search game
[2] By no means did Bing have exactly the same search results as Google
[3] One of the worst things about Yahoo changing over to Bings results last year was that in the US and in many countries around the world we were suddenly down to only two search voices Googles and Bings
[4] However the increases were indicative that Bing had made some change to its search algorithm which was causing its results to be more Google-like
[5] These searches returned no matches on Google or …</td><td>[1] We will also use this information to help improve our products and services
[2] Again I've bolded the key parts
[3] For 15 years I've covered search
[4] Google has run a sting operation that it says proves Bing has been watching what people search for on Google the sites they select from Google's results then uses that information to improve Bings own search listings
[5] By no means did Bing have exactly the same search results as Google
[6] One of the worst things about Yahoo changing over to Bings results last year was that in the US and in many countries around the world ..</td></tr>
<tr><th>EXP
Recall ROUGE-1: 0.403</th><th>T10.EXP
Recall ROUGE-1: 0.438</th></tr>
<tr><td>[1] If it's not illegal is what Bing may be doing unfair somehow cheating at the search game.
[2] By no means did Bing have exactly the same search results as Google.
[3] One of the worst things about Yahoo changing over to Bings results last year was that in the US and in many countries around the world we we're suddenly down to only two search voices: Google's and Bing's.
[4] However the increases we're indicative that Bing had made some change to its search algorithm which was causing it's results to be more Google-like.
[5] These searches returned no matches on Google or...</td><td>[1] If it's not illegal is what Bing may be doing unfair somehow cheating at the search game.
[2] I don't know how else to call it but plain and simple cheating.
[3] Is it Cheating?
[4] If they started to appeared at Bing after Google that would mean that Bing took Google's bait and copied its results.
[5] Now Google began to strongly suspect that Bing might be somehow copying it's results in particular by watching what people we're searching for at Google.
[6] Google says it doesn't know why they didn't all work but even having a few appear was enough to convince the company that Bing was copying it's results ..</td></tr>
</table>

Figure 4.12: Examples for DocID 2 (208 sentences and 80 tweets)

We noticed that the LOCAL and LD.EXP summaries contained the most sentences with the word '*Google*' and '*Bing*' and also extracted the same sentence ('*Google has run a sting operation that it says proves Bing has been watching what people search for on Google the sites they select from Google's results then uses that information to improve Bing's own search listings'*), which is also included in the reference summaries.

All the tweet-biased summaries (EXP, EXP_S, and T10.EXP) were found to incorporate the tweets in the sentence selection. Most of the tweets are on the title of the documents, which contains the word '*BING*' and '*cheating*', and this is reflected in the sentences extracted by the tweet-biased summarization system. Note that EXP and EXP_S extracted the same sentences, which suggested that document-tweets similarity and sentence-tweets similarity may not have significant effect on the Affinity Graph algorithm. However, T10.EXP generated a different but better summary (and a higher Recall ROUGE score) compared to EXP and EXP_S. This indicated that using fewer tweets and important tweets might be enough to generate a good summary. Also, note that the three Affinity Graph settings extracted the same first sentence ('*If it's not illegal is what Bing may be doing unfair somehow cheating at the search game*').

A different summary analysis was seen in DocID311 (Figure 4.13). Perhaps because this is a long document, we see that a different topic was selected in the reference summaries; one summary describes Irish debt and another Irish house prices. We believe that due to the length of the document, it's harder for the human summarizer to choose the best sentence to best represent the document.

In the LOCAL and LD.EXP summaries, information from the Local Document improved the *if_score* of the sentences, where both summaries extracted the same three sentences. We can also see that all summaries that were generated using different Expanded tweets settings have different contents. We assumed that because the number of tweets is low, the Affinity Graph summariser could not identify the relevant topic to extract the best sentences for the summary. This may also have caused the low ROUGE score, but since the difference is not significant, we could not conclude which setting produced the best summaries.

DOCUMENT (DocID 311)
TITLE: When Irish Eyes Are Crying

BASELINE 1
Recall ROUGE-1: 0.493

First Iceland. Then Greece. Now Ireland, which headed for bankruptcy with its own mysterious logic. In 2000, suddenly among the richest people in Europe, the Irish decided to buy their country from one another. After which their banks and government really screwed them. So where's the rage? When I flew to Dublin in early November, the Irish government was busy helping the Irish people come to terms with their loss. It had been two years since a handful of Irish politicians and bankers decided to guarantee all the debts of the country's biggest banks, but the people were only now…

BASELINE 2
Recall ROUGE-1: 0.337

First Iceland. When I flew to Dublin in early November, the Irish government was busy helping the Irish people come to terms with their loss. The two other big Irish banks, Bank of Ireland and, especially, Allied Irish Banks (A.I.B.), remained Ireland's dirty little secrets. Even in an era when capitalists went out of their way to destroy capitalism, the Irish bankers set some kind of record for destruction. Ireland's financial disaster shared some things with Iceland's. In recognition of the spectacular losses, the entire Irish economy has almost dutifully collapsed. Yet when I arrived, in early November 2010, Irish …

TWEET EXAMPLES

A new #longreads from Michael Lewis in March's @vanityfair . When Irish Eyes Are Crying about Irish bankers.
RT @DylanRatigan Talking w/ Michael Lewis of the Big Short. Why no protests in Ireland? Lewis's piece: / #Banksters
RT: @DylanRatigan Talking with Michael Lewis of the Big Short. Why no protests in Ireland? Here's Lewis's piece: #msnbc
@vanityfairmag Not even the godlike Michael Lewis can justify another "Irish Eyes" headline. For shame!
Excellent article on the Irish banking collapse by Michael Lewis (author of The Big Short).
RT @stunoble : Stunning piece on the the irish financial crisis: When Irish Eyes Are Crying -
Must-read Michael Lewis piece on how the Irish discovered optimism just in time for it to bury them: @VanityFairMag :
Good article about the Irish economic situation: When Irish Eyes Are Crying: vanityfair.com: via @addthis
Michael Lewis ->; national treasure. : #VanityFair #Ireland
When Irish Eyes Are Crying | Business | Vanity Fair <— A long but interesting read.
Continuing Lewis' narrative tour of financial collapse:

REFERENCE (MANUAL) SUMMARIES

Summary 1	Summary 2
[1] In 2000, suddenly among the richest people in Europe, the Irish decided to buy their country from one another. [2] It had been two years since a handful of Irish politicians and bankers decided to guarantee all the debts of the country's biggest banks, but the people were only now getting their minds around what that meant for them. [3] As the sum total of loans made by Anglo Irish, most of it to Irish property developers, was only 72 billion euros, the bank had lost nearly half of every dollar it invested. [4] Ireland's financial disaster shared some things with Iceland's.	[1] Now Ireland, which headed for bankruptcy with its own mysterious logic. [2] An Irish economist named Morgan Kelly, whose estimates of Irish bank losses have been the most prescient, made a back-of-the-envelope calculation that puts the losses of all Irish banks at roughly 106 billion euros. [3] Kelly saw house prices rising madly and heard young men in Irish finance to whom he had recently taught economics try to explain why the boom didn't trouble them. [4] The moment people cease to believe that house prices will rise forever, they will notice what a terrible long-term investment real estate has become and flee the market, and the market will crash.

LOCAL
Recall ROUGE-1: 0.317

[1] That time was before the Irish government used ECB money to pay off the foreign bondholders in Irish banks
[2] The two other big Irish banks Bank of Ireland and especially Allied Irish Banks AIB remained Irelands dirty little secrets
[3] The Irish bank losses have obviously bankrupted Ireland but the Irish finance minister does not want to talk about that
[4] AIB lent the money for 6 of the 15 Anglo Irish for just 1 as a colender with AIB On Irish national radio recently the insolvent property developer Simon Kelly whose family's real-estate portfolio has run up bad debts of 2 billion euros confessed that ..

EXP_S **Recall ROUGE-1: 0.317**	**LOCAL.EXP** **Recall ROUGE-1: 0.346**
[1] The Irish bank losses have obviously bankrupted Ireland but the Irish finance minister does not want to talk about that [2] Two weeks later Lenihan will be compelled by the European Union to invite the IMF into Ireland relinquish control of Irish finances and accept a bailout package [3] Ask Irish property developers who they imagined was going to live in the Irish countryside and they all laugh the same uneasy laugh and offer up the same list of prospects Poles foreigners looking for second homes entire departments of Irish government workers who would be shipped to the sticks in a …	[1] That time was before the Irish government used ECB money to pay off the foreign bondholders in Irish banks [2] After all the vast majority of the construction was being funded by Irish banks [3] The Irish bank losses have obviously bankrupted Ireland but the Irish finance minister does not want to talk about that [4] AIB lent the money for 6 of the 15 Anglo Irish for just 1 as a colender with AIB [5] On Irish national radio recently the insolvent property developer Simon Kelly whose family's real-estate portfolio has run up bad debts of 2 billion euros confessed that the only time…
EXP **Recall ROUGE-1: 0.268**	**Top 10 EXP** **Recall ROUGE-1: 0.307**
[1] Anyone who has been anywhere near an Irish Catholic family knows the member who has had the most recent run of bad luck enjoys exalted status the right to do pretty much whatever he wants while everyone else squirms in silence [2] Underlying the public opinion polls that show the Irish feel a lot better about the minister of finance than they do about other politicians in his party is a common unspoken understanding of his bravery [3] In America the banks went down but the big shots in them still got rich in Ireland the big shots went down with …	[1] In America the banks went down but the big shots in them still got rich in Ireland the big shots went down with the banks [2] Four different Irish people told me on great authority that Cowen had faxed Irelands 440 billion euro bank guarantee into the European Central Bank from a pub [3] Lehman Brothers had failed two days earlier shares of Irish banks were plummeting and big corporations were withdrawing their deposits from them [4] In September 2010 the last big chunk of money the Irish banks owed the bondholders 26 billion euros came due

Figure 4.13: Examples for DocID 311 (694 sentences and 11 tweets)

The next analysis is for documents with few sentences (Figure 4.14).

<table>
<tr><th colspan="2">DOCUMENT (DocID 55)
TITLE: Would-Be Suicide Bomber Killed by Unexpected SMS From Mobile Carrier</th></tr>
<tr><th colspan="2">BASELINE 1
Recall ROUGE-1: 0.758</th></tr>
<tr><td colspan="2">An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve, inadvertently thwarting a planned attack on revelers in Moscow, according to The Daily Telegraph. The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square, a plan that was foiled when her wireless carrier sent her an SMS while she was still at a safe house, setting off the bomb and killing her. The message reportedly wished her a Happy New Years, according to the report, which sourced the info from security forces in Russia.</td></tr>
<tr><th colspan="2">BASELINE 2
Recall ROUGE-1: 0.745</th></tr>
<tr><td colspan="2">An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bomber's vest bomb in Russia New Year's Eve, inadvertently thwarting a planned attack on revelers in Moscow, according to The Daily Telegraph. The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square, a plan that was foiled when her wireless carrier sent her an SMS while she was still at a safe house, setting off the bomb and killing her. If true, the SMS might be the only time that a wireless carrier's SMS message has ever been useful. The …</td></tr>
<tr><th colspan="2">TWEET EXAMPLES</th></tr>
<tr><td colspan="2">Schadenfreude Alert! RT @JonHenke : This. Is. Awesome. RT @sorendayton Story of the day ...
RT @JonHenke : This. Is. Awesome. RT @sorendayton Story of the day ...
[feed] Unexpected SMS Kills Would-Be Suicide Bomber: A suicide bomber plotting to kill Russians celebrating New ...
Would-Be Suicide Bomber Killed by Unexpected SMS From Mobile Carrier | Threat Level | Wired.com
Happy New Years msg kills: SMS Kills Would-Be Suicide Bomber: A suicide bomber in Russia...
Bizarre. Would-be suicide bomber killed by SMS from mobile carrier:
RT @DaveMedlo : Suicide bomber killed by service provider text >> PMSL!! :o)
Would-Be Suicide Bomber Killed by Unexpected SMS from Mobile Carrier >> SPAM saves lives
RT @wired : Would-Be Suicide Bomber Killed by Unexpected SMS from Mobile Carrier >> SPAM saves lives
Blimey. RT @wired : Would-Be Suicide Bomber Killed by Unexpected SMS from Mobile Carrier >> SPAM saves lives
Would-Be Suicide Bomber Killed by Unexpected SMS from Mobile Carrier >> SPAM saves lives
Heh. "If true, the SMS might be the only time that a wireless carrier's SMS message has ever been useful." -
Would-Be Suicide Bomber Killed by Unexpected SMS From Mobile Carrier...
"Would-Be Suicide Bomber Killed by Unexpected SMS From Mobile Carrier" -- "[It] reportedly wished her a Happy New Years."
Unexpected SMS Kills Would-Be Suicide Bomber: A suicide bomber plotting to kill Russians celebrating New Year's ...
Unexpected SMS Kills Would-Be Suicide Bomber - "If true, the SMS might be the only time that a wireless carrier's SMS message has ever been useful."
RT @kleptones : Worthy of Chris Morris: "Would-Be Suicide Bomber Killed by Unexpected SMS From Mobile Carrier"
Unexpected SMS Kills Would-Be Suicide Bomber: A suicide bomber plotting to kill Russians celebrating New Year's ...</td></tr>
<tr><th colspan="2">REFERENCE SUMMARIES</th></tr>
<tr><th>Summary 1</th><th>Summary 2</th></tr>
<tr><td>[1] An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve, inadvertently thwarting a planned attack on revelers in Moscow, according to The Daily Telegraph.
[2] The message reportedly wished her a Happy New Years, according to the report, which sourced the info from security forces in Russia.
[3] Cell phones are often used as makeshift detonators by terrorist and insurgent groups.
[4] If true, the SMS might be the only time that a wireless carriers SMS message has ever been useful.</td><td>[1] The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square, a plan that was foiled when her wireless carrier sent her an SMS while she was still at a safe house, setting off the bomb and killing her.
[2] Cell phones are often used as makeshift detonators by terrorist and insurgent groups.</td></tr>
<tr><td colspan="2"></td></tr>
<tr><th colspan="2">Local Document
Recall ROUGE-1: 0.617</th></tr>
<tr><td colspan="2">[1] If true the SMS might be the only time that a wireless carriers SMS message has ever been useful
[2] The authorities suspect the female bomber was part of the same Jihadist group that is suspected of hitting Moscow's airport on Monday with a suicide bomb attack that killed 35
[3] The message reportedly wished her a Happy New Years according to the report which sourced the info from security forces in Russia
[4] Cell phones are often used as makeshift detonators by terrorist and insurgent groups</td></tr>
<tr><th>EXP$_S$
Recall ROUGE-1: 0.611</th><th>Local Document+EXP
Recall ROUGE-1: 0.718</th></tr>
<tr><td>[1] The authorities suspect the female bomber was part of the same Jihadist group that is suspected of hitting Moscow's airport on Monday with a suicide bomb attack that killed 35
[2] An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve inadvertently thwarting a planned attack on revelers in Moscow according to The Daily Telegraph
[3] The would be suicide bomber was planning to detonate a suicide belt bomb near Red Square a plan that was foiled when her wireless carrier sent her an SMS while she was still ..</td><td>[1] If true the SMS might be the only time that a wireless carriers SMS message has ever been useful
[2] An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve inadvertently thwarting a planned attack on revelers in Moscow according to The Daily Telegraph
[3] The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square a plan that was foiled when her wireless carrier sent her an SMS while she was still at a safe house setting off the bomb and killing her</td></tr>
<tr><th>EXP
Recall ROUGE-1: 0.745</th><th>Top 10 EXP
Recall ROUGE-1: 0.745</th></tr>
<tr><td>[1] If true the SMS might be the only time that a wireless carriers SMS message has ever been useful
[2] The message reportedly wished her a Happy New Years according to the report which sourced the info from security forces in Russia
[3] An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve inadvertently thwarting a planned attack on revelers in Moscow according to The Daily Telegraph
[4] The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square a plan that was foiled when her …</td><td>[1] If true the SMS might be the only time that a wireless carriers SMS message has ever been useful
[2] The message reportedly wished her a Happy New Years according to the report which sourced the info from security forces in Russia
[3] An unexpected and unwanted text message from a wireless company prematurely exploded a would-be suicide bombers vest bomb in Russia New Years Eve inadvertently thwarting a planned attack on revelers in Moscow according to The Daily Telegraph
[4] The would-be suicide bomber was planning to detonate a suicide belt bomb near Red Square a plan that was foiled when her …</td></tr>
</table>

Figure 4.14: Examples for Doc ID 55 (6 sentences and 21 tweets)

For DocID 55, the document content is straightforward, thus the reference summaries have the same topic, with one sentence extracted by the human summarizer in common.

All Expanded Tweets summaries have at least one sentence in common, where it included the main topic from the tweets ('*The would-be suicide bomber*'). The sentences from the different summaries are also found in the reference summaries, resulting in a high Recall ROUGE score. The EXP and T10.EXP setting also produced exactly the same summary. Thus, for documents with short length, it seems that the number of tweets used to extract sentences does not matter.

The same analysis is shown in DocID 426 (Figure 4.15), where the EXP and T10.EXP have the same summaries. We can see that almost all of the tweets have the title of the document, thus the word 'Disney', 'Fox', and 'Hulu' appear the most. This may give greater influence in sentence selection for summaries generated using Affinity Graph, since at least one same sentence is extracted and appears in the different summary. The human summarizer also extracts the same sentence in the reference summary.

We can see that most of the tweets are taken from the document's title or the first few words from the first paragraph. For the documents with a high number of tweets (Doc ID 2 and Doc ID 55), we can see that RTs (ReTweets) of the first tweet from the original tweet's user also dominated the tweet collection. Only a few tweets have personal opinion or information regarding the document. We also agree with (Sharifi et al., 2010) that longer tweets do not always represent the main ideas of the document and/or tweets, but contain more "emotional" comments on a topic.

However, we believe the number of tweets related to the document is beneficial for longer documents. This was showed in Figure 4.13 where we can see that summary generated by EXP and T10.EXP have different sentences, but are discussing the same topic. For long documents but with a low number of tweets, the related tweets do not help much, because the summarizer could not identify the main topic for the document. Hence, all of the expanded summaries have different content and sentences.

<table>
<tr><th colspan="2">DOCUMENT (Doc ID 426)
TITLE: Disney & Fox Consider Pulling Content From Hulu [REPORT]</th></tr>
<tr><th colspan="2">BASELINE 1
Recall ROUGE-1: 0.725</th></tr>
<tr><td colspan="2">Uncertainty about Hulu's business model has prompted some of its media backers to contemplate pulling content to run elsewhere, according to a report. The Wall Street Journal reports today that NBC Universal, News Corp. and Walt Disney Co. are "increasingly at odds" over Hulu's business model and are worried that running content on the site is endangering their own businesses. (A subscription is required to access the link.) As a result, Disney and Fox Broadcasting owner News Corp. are considering pulling content from Hulu and are "moving to sell more programs to Hulu competitors that deliver television over the Internet.</td></tr>
<tr><th colspan="2">BASELINE 2
Recall ROUGE-1: 0.761</th></tr>
<tr><td colspan="2">Uncertainty about Hulu's business model has prompted some of its media backers to contemplate pulling content to run elsewhere, according to a report. The Wall Street Journal reports today that NBC Universal, News Corp. and Walt Disney Co. are "increasingly at odds" over Hulu's business model and are worried that running content on the site is endangering their own businesses. As a result, Disney and Fox Broadcasting owner News Corp. are considering pulling content from Hulu and are "moving to sell more programs to Hulu competitors that deliver television over the Internet, including Netflix, Microsoft and Apple," according to the …</td></tr>
<tr><th colspan="2">TWEET EXAMPLES</th></tr>
<tr><td colspan="2">Disney & Fox Consider Pulling Content From Hulu [REPORT] -
Disney & Fox Consider Pulling Content From Hulu [REPORT] via @mashablemedia @mashable I don’t support this idea from Fox
Disney & Fox Consider Pulling Content From Hulu [REPORT] via @mashablemedia @mashable
#mashable Disney & Fox Consider Pulling Content From Hulu [REPORT]
Disney & Fox Consider Pulling Content From Hulu [REPORT] via @mashablemedia @mashable
#Disney & #Fox Consider Pulling Content From Hulu [REPORT] via @mashable #media
Say it is'nt so! RT @agripundit : Bad News. Disney & Fox Consider Pulling Content From Hulu [REPORT
Disney & Fox Consider Pulling Content From Hulu [REPORT]
Disney & Fox Consider Pulling Content From Hulu [REPORT]
I hope they put it on Netflix. RT @MovieViral : Disney & Fox Consider Pulling Content From Hulu [REPORT] via @mashable</td></tr>
<tr><th colspan="2">REFERENCE SUMMARIES</th></tr>
<tr><th>Summary 1</th><th>Summary 2</th></tr>
<tr><td>[1] Uncertainty about Hulu's business model has prompted some of its media backers to contemplate pulling content to run elsewhere, according to a report.
[2] Created in 2007, Hulu was designed to let News Corp. and its other media backers offset the influence of YouTube and pirated versions of TV shows on the Internet.
[3] But since 2008, sales execs at Fox and NBC have complained that the site is drawing viewers from Fox.com and NBC.com, respectively.</td><td>[1] Uncertainty about Hulu's business model has prompted some of its media backers to contemplate pulling content to run elsewhere, according to a report.
[2] As a result, Disney and Fox Broadcasting owner News Corp. are considering pulling content from Hulu and are 'moving to sell more programs to Hulu competitors that deliver television over the Internet, including Netflix, Microsoft and Apple,' according to the article.</td></tr>
<tr><td colspan="2"></td></tr>
<tr><th colspan="2">Local Document
Recall ROUGE-1: 0.683</th></tr>
<tr><td colspan="2">[1] The story also reports that Hulu management has discussed recasting Hulu as an online cable operator that would use the web to send live TV channels and video-on-demand content to subscribers
[2] As a result Disney and Fox Broadcasting owner News Corp are considering pulling content from Hulu and are moving to sell more programs to Hulu competitors that deliver television over the Internet including Netflix Microsoft and Apple according to the article
[3] Created in 2007 Hulu was designed to let News Corp and its other media backers offset the influence of YouTube and pirated versions of TV shows on the Internet</td></tr>
<tr><th>EXP$_S$
Recall ROUGE-1: 0.704</th><th>Local Document+EXP
Recall ROUGE-1: 0.746</th></tr>
<tr><td>[1] But since 2008 sales execs at Fox and NBC have complained that the site is drawing viewers from Fox.com and NBC.com respectively
[2] Uncertainty about Hulu’s business model has prompted some of its media backers to contemplate pulling content to run elsewhere according to a report
[3] The story also reports that Hulu management has discussed recasting Hulu as an online cable operator that would use the web to send live TV channels and video-on-demand content to subscribers
[4] Hulu reps could not be reached for comment about the report
[5] As a result Disney and Fox Broadcasting owner News Corp are considering pulling …</td><td>[1] The Wall Street Journal reports today that NBC Universal News Corp and Walt Disney Co are increasingly at odds over Hulu’s business model and are worried that running content on the site is endangering their own businesses
[2] Uncertainty about Hulu’s business model has prompted some of its media backers to contemplate pulling content to run elsewhere according to a report
[3] As a result Disney and Fox Broadcasting owner News Corp are considering pulling content from Hulu and are moving to sell more programs to Hulu competitors that deliver television over the Internet including Netflix Microsoft and Apple according to the article</td></tr>
<tr><th>EXP
Recall ROUGE-1: 0.697</th><th>Top 10 EXP
Recall ROUGE-1: 0.697</th></tr>
<tr><td>[1] The Wall Street Journal reports today that NBC Universal News Corp and Walt Disney Co are increasingly at odds over Hulu’s business model and are worried that running content on the site is endangering their own businesses
[2] As a result Disney and Fox Broadcasting owner News Corp are considering pulling content from Hulu and are moving to sell more programs to Hulu competitors that deliver television over the Internet including Netflix Microsoft and Apple according to the article
[3] Created in 2007 Hulu was designed to let News Corp and its other media backers offset the influence of YouTube and pirated …</td><td>[1] The Wall Street Journal reports today that NBC Universal News Corp and Walt Disney Co are increasingly at odds over Hulus business model and are worried that running content on the site is endangering their own businesses
[2] As a result Disney and Fox Broadcasting owner News Corp are considering pulling content from Hulu and are moving to sell more programs to Hulu competitors that deliver television over the Internet including Netflix Microsoft and Apple according to the article
[3] Created in 2007 Hulu was designed to let News Corp and its other media backers offset the influence of YouTube and pirated versions</td></tr>
</table>

Figure 4.15: Examples for Doc ID 426 (9 sentences and 10 tweets)

We can also see that there is a range of scores for the different summaries and also summaries with the same score. However, we can also see that there are summaries with the same Recall ROUGE-1 score that have different sentences, such as in LOCAL and LD.EXP (Doc ID 2) and

LOCAL and EXP_s (Doc ID 311). Both documents have more than 200 sentences, which means more sentences to choose from, and therefore a greater variety of possible summaries that could be generated.

It is apparent from these figures that the Affinity Graph approach was able to generate comparable summaries using the related tweets. Even though the ROUGE score did not show significant difference between the tweet-biased Affinity Graph settings, some of the summaries extracted different sentences, resulting in different summary content. It appears that longer documents need related tweets more – compared to shorter documents – to help the Affinity Graph find a certain topic to generate its summary. Without the related tweets, the Affinity Graph would create a summary with mix topics of the documents, as shown in Figure 4.13.

4.4.2 ROUGE Score Correlation

We were interested to further examine the relationship between the number of sentences/tweets and the Recall rouge-1 scores. We analysed the correlation between the number of sentences in a document and the Recall ROUGE-1 scores for each setting (Figure 4.16).

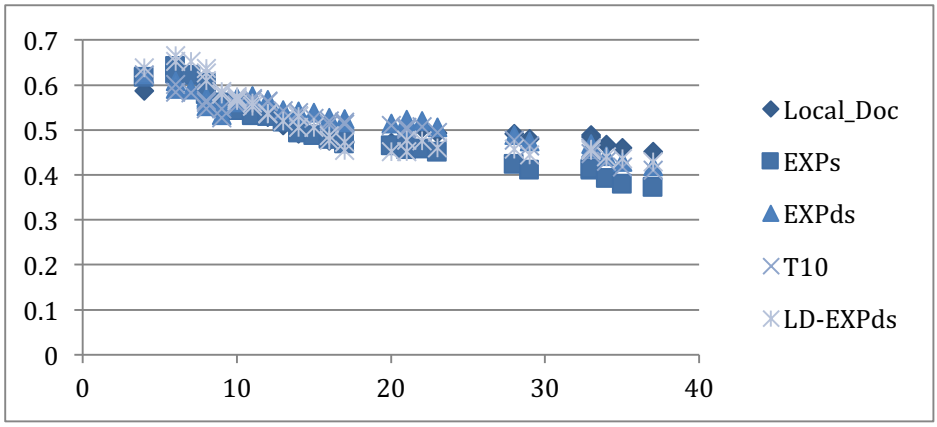


Figure 4.16: Number of Sentences vs Recall ROUGE-1 scores

Figure 4.16 shows the rolling average for 20 documents plotting the number of sentences against ROUGE-1 scores. We can see that the correlation (average r=-0.9) has the same pattern with

Figure 3.18, where the ROUGE-1 scores decrease with longer documents. We believed that documents with fewer words had fewer choices of sentences, thus much easier to get higher ROUGE scores.

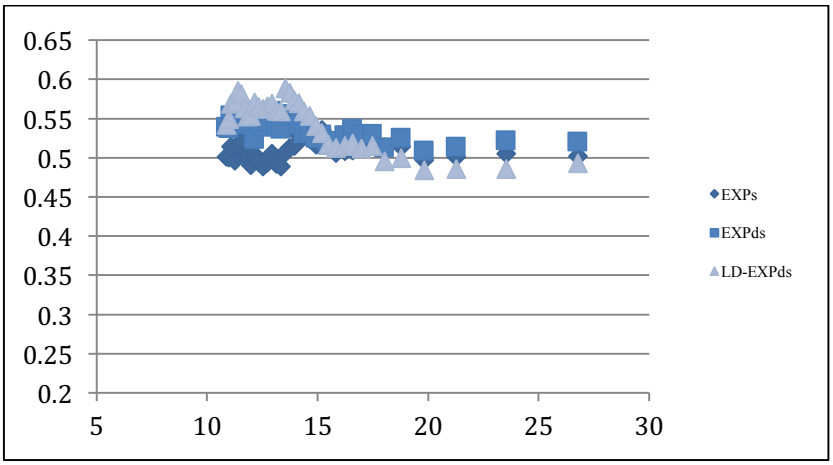


Figure 4.17: Number of Tweets vs. Recall ROUGE-1 scores

We also examined the effect of the number of related tweets on ROUGE-1 (Figure 4.17). We removed the results of T10.EXP and LOCAL because the settings use either a fixed number or no tweets. No correlation was found ($r < 0.1$ and $p > 0.05$) between the number of tweets and the Recall ROUGE scores. The number of tweets does not appear to have any effect on ROUGE.

4.4.3 Expanded Tweets (EXP) ROUGE Score Analysis

In the EXP setting, we were also interested to see if a different number of tweets had any effect on the summary generated. For this, we experimented using the Top 10 tweets based on the document-tweet similarity value (Part 1a in Figure 4.2). We chose only to test with Top 10 tweets because it was the minimum number of tweets for the local document and to ensure all documents are included in the experiment. In Table 4.5, we can see that EXP_s and T10.EXP shows no notable difference in ROUGE. T10.EXP performed equally well compared with Local Document summaries. This is shown in Figure 4.18.

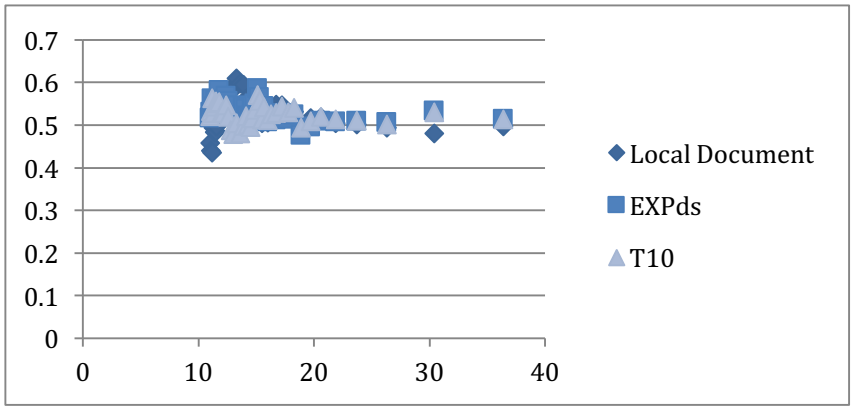


Figure 4.18: Recall ROUGE-1 scores sorted by number of tweets for Local Document, Expanded Tweet and EXP -Top10

We can see that there is not much difference in the Recall ROUGE-1 scores for EXP and T10.EXP. In Figure 4.18, we excluded the results for ten tweets because we are only interested analyzing the scores when the original number of tweets are more than ten. We can see that for the document with more tweets, their Recall ROUGE-1 scores are almost similar.

4.5 Discussion

In this chapter, we discussed the application the Affinity Graph algorithm to a new dataset (Tweet-WebDoc) and used a different type of document (tweets) to generate a summary. For this work, we found similarities with work by P. Hu, Ji, et al. (2011b) and P. Hu, Sun, et al. (2011), who focused on building a social context summarization using user tags in a social bookmarking website. Both works assume that constraint information (i.e. tagging and tweets) would help summarisers identify more relevant information and thus improve summary accuracy.

We tested the Affinity Graph algorithm by defining related neighbourhood setting as EXP, EXP_s, and T10_EXP. We believed that the document similarity value ($sim_{doc-tw}(d_i,t_j)$) and the sentence similarity value ($sim_{sen-tw}(s_i,t_j)$) could create a stronger relationship between the summary and its tweets. Figure 4.10 and Figure 4.11, shows that the social content of the local document is more useful when it is being used together with the local document.

To test our work, we created a new summary evaluation dataset based on documents and tweets gathered by Yulianti et al. (2015). Our ROUGE results revealed that the tweet-biased summarization using the Affinity Graph did not show significance difference when compared with Local Document summaries. Only EXP was significantly better when compared with EXP_S in its Precision score. The same result was shown in the comparison with TBS, where the ROUGE score did not show significant differences in the results.

We performed a post-hoc power analysis of the Recall Rouge-1 scores, showing low power scores (0.1-0.2) for all paired settings. We found that we would need a large number of summarized documents (300-8000), in order to identify a significant difference between settings. If there are significant differences in the ROUGE scores, they are not visible due to the small number of summarized documents we have in our dataset.

In the work of P. Hu, Ji, et al. (2011b) and P. Hu, Sun, et al. (2011), they created a new dataset by downloading 200 bookmarked CNN news articles via a social-tagging website[8]. Not much discussion was found in their paper on how they developed their dataset but they mentioned that they extracted 2186 tags for the 200 documents from 1194 users (P. Hu, Sun, et al., 2011). They reported that they produced significant results when compared to a baseline system, but did not mention if any power analysis was conducted. They show that the use of social media could improve document summarization by adding extra information to an Affinity Graph.

We conducted an analysis to see if the summaries generated by the Affinity Graph were better than the LOCAL summaries despite having similar ROUGE scores, a so-called sentence granularity problem (Nenkova & McKeown, 2011). We found that summaries were different, thus this has encouraged us to explore other evaluation methods for document summarization.

An alternative summary evaluation approach was tried by Mackie et al. (2014) and Yulianti et al. (2015), who explored the use of crowdsourcing to evaluate microblog summaries, see also (Lloret, Plaza, & Aker, 2013; Mackie et al., 2014). A study by Wang, Zhu, Li, Chi, & Gong (2011) examined

[8] https://delicious.com/

user studies to test datasets and algorithms. Based on this, we are interested to discover if measurement of summaries based on user preference might produce more powerful results.

4.6 Conclusion and What's Next?

We used an Affinity Graph for single document summaries using social media content and contributed the following:

1. The use of the Affinity Graph to integrate social media content was tried in four settings (EXP, EXP_S, T10.EXP, and LD.EXP).
2. We enhanced an existing dataset (Tweet-WebDoc dataset) with manually generated summaries.
3. We developed a summary sentence selection system SESys.
4. We found that the Affinity Graph approach was able to extract and create different summaries compared to the LOCAL summaries, despite small differences in their ROUGE scores.
5. In the ROUGE-document length correlation, we found a similar result as discussed in Chapter 3: there is a correlation between the number of sentences and the ROUGE score. We also discovered that the number of related tweets has no effect on the ROUGE scores.

We identified new questions to further analyse the summaries generated from our summarization system:

- We will conduct a study of human summary preferences;
- We will study what are the aspects of a summary that people consider when choosing one over another.

These experiments are described in Chapter 5.

Chapter 5
Summary Evaluation: Relevant vs. Judgement

Our findings in Chapter 4 discussed a tweet-biased Affinity Graph approach. However, few significant differences were shown in the ROUGE scores. A power analysis suggested this was due to the small dataset used in our experiments. A manual examination of the summaries indicated that summaries of long documents with many related tweets were more topic-focused. We believed that the approach was able to produce better summaries, but we needed a different evaluation method.

Previous studies on document summarization primarily concentrated on automated evaluation of summaries. In document summarization, there have been relatively few studies on human judgement on summaries (Lloret et al., 2013; Mackie et al., 2014). We believed that by applying a human judgement evaluation approach, we could draw more conclusive finding.

We identified the third research question (RQ3) for the thesis:

"*Is a crowdsourced human judgement approach a better evaluation compared to the standard automated summary evaluation?*"

We also identified a fourth research question (RQ4):

"*Will a tweet-biased Affinity Graph approach be preferred over LOCAL settings?"*

In this chapter, we will discuss the following contributions:

1. The use of a crowdsourcing platform to evaluate summaries. Here, we asked people to judge which summary is the best to represent the document and explain their reason for preferring one summary over another. Discussion on the crowdsourcing platform setups is included.
2. Analysis of human judgements for the different tweet-biased Affinity Graph settings.
3. We also discuss the qualitative findings on the features of the preferred summaries.

We defined the following terms used in our experiments which, will be used throughout this chapter:

- Jobs: a summary judgement 'task' which we released to Crowdsource workers.
- Participants: the crowdsource workers that took part in our released jobs.

5.1 Background Work

Crowdsourcing is the process of getting work (services, ideas, or content) from an online community. It is ideal for large-scale, repetitive tasks which require a scalable workforce in order to get the task done in a short time. Crowdsourcing services have been used in information retrieval studies. For example, Amazon Mechanical Turk[1] and CrowdFlower[2] have been used in relevance evaluation (Alonso et al., 2008; Grady & Lease, 2010; Hosseini, Cox, Milić-Frayling, Kazai, & Vinay, 2012), video/text annotation (Finin et al., 2010; Nowak & Rüger, 2010; S. Park, Mohammadi, Artstein, & Morency, 2012; Snow, O'Connor, Jurafsky, & Ng, 2008) and user studies (Kittur, Chi, & Suh, 2008; Komarov, Reinecke, & Gajos, 2013). We found that Štajner et al. (2013) used similar crowdsourcing evaluations in their work, but not specifically to summarize the tweets.

[1] https://www.mturk.com/mturk/welcome

[2] http://www.crowdflower.com

5.1.1 Use of Crowdsource Platform to Evaluate Summaries

Most work on the document summarization using crowdsourcing focuses on creating summaries (Lloret et al., 2013), where the capability and reliability of crowdsourcing were discussed. However, some past work used crowdsourcing for evaluation (Mackie et al., 2014; Yulianti et al., 2015).

Mackie et al. (2014) discussed different evaluation measures for document summarization that included automatic evaluation (using ROUGE, Jensen-Shannon Divergence, and Fraction of Topic Words) and participant preference (pair-wise evaluation using CrowdFlower) for microblog summarization. They applied three different systems (namely Centroid, SumBasic and Hybrid) to summarize tweets from four microblog datasets, which includes the TREC 2011 Microblog track dataset.

Pairwise evaluation was used in an earlier study examining a limited number of summarization systems (Yulianti et al., 2015). In addition to participant preferences, they also gathered feedback from participants on why summaries were preferred.

There has been little past work on evaluating a proposed document summarization using both ROUGE-based and Crowdsource evaluations. As discussed in Chapter 4, we developed a summarization dataset. By evaluating the summaries in two different approaches, we believed that our work would contribute to the study of implementing human judgement as an alternative evaluation for document summarization.

5.2 Experiment Setup

For our experiments, we have chosen the CrowdFlower online service. In this section, we discuss the experiment setup for the Affinity Graph and the CrowdFlower settings. We also discuss the pilot tests done prior to the real experiments and also the test questions that serve as a quality control for the CrowdFlower jobs.

5.2.1 Affinity Graph Settings for CrowdFlower

To design our experiments, we prepared pairs of summaries from the same document of different Affinity Graph settings and asked participants to judge the summary. The main reason that we asked the participants to select from pairs of summaries (rather than choosing more than 2 summaries or give scores to the summaries) as Jones, Brun, & Boyer (2011) found that the participants are able to make reliable decisions when asked to compare rather than rate. We agreed with Lloret et al. (2013) that the task be as simple as possible for participants. Pairwise comparison was also a preferred method in work by Diakopoulos, Choudhury, & Naaman (2012), Glaser & Schütze (2012), Sanderson, Paramita, Clough, & Kanoulas (2010) and Yang et al. (2011).

We chose paired summaries based on neighborhood settings. EXP used the maximum available similarity information: document-tweet and sentence-tweet. Therefore, we paired EXP with all the other settings (LOCAL, EXPS, T10.EXP, and LD.EXP). We were also interested to know if LOCAL summaries would be chosen compared to the other setting that uses social content. Thus, we paired it together with EXP_S and LD.EXP. We also compared EXP and the TBS_{sn} because both the summaries used similar information, but in different ways. The chosen pairs were loaded into seven jobs:

1) EXP and LOCAL
2) EXP and EXP_S
3) EXP and T10.EXP
4) EXP and LD.EXP
5) EXP_S and LOCAL
6) LD.EXP and LOCAL
7) EXP and TBS_{sn}

5.2.2 Test Questions

We need to create Test Questions as a mechanism for quality control. These are questions with known answers used to test participants' accuracy. These test questions help ensure only the answers from competent participants are included in the job results.

Each test question was set up as a judgement question, where a summary from the target document was shown next to a summary, from a totally different document. We created 123 test questions for each job, one for every four judgments. The test question is placed in a random position on each page. In the job's Data section, we have set the test questions as TRUE in the _golden column, to indicate that this row is the test question as shown in Figure 5.1.

Judgments	Agreement	_golden	id	summary_one	summary_two	text	title	which_summary_do_yo...	which_summary_do_yo...
15		TRUE	1_gold	But Lundgren's commen...	"Some of them lost their...	West Palm Beach, Flori...	'Brilliant Bus' shrinking d...	Summary 2	Summary 1 was genera...
13		TRUE	2_gold	If you say so, I said. "La...	(CNN) -- "Warmly welco...	Editor's note: Mike Dow...	'L' is for losers in L.A. sp...	Summary 1	Summary 2 was genera...
13		TRUE	3_gold	"The field of substance ...	They weren't necessaril...	(CNN) -- Last year we p...	10 classic American exp...	Summary 2	Summary 1 was genera...
10		TRUE	4_gold	" Got a question about e...	Jakarta, Indonesia (CNN...	Jakarta, Indonesia (CNN...	28 bodies recovered aft...	Summary 2	Summary 1 was genera...
9		TRUE	5_gold	Tornadoes kill 70 people...	As well as smuggled bo...	World-renowned chef, a...	8 things to know before ...	Summary 2	Summary 1 was genera...
17		TRUE	6_gold	This will be simply impo...	They're nutritious, eco-fr...	(CNN) -- According to a ...	A traveler's guide to eati...	Summary 2	Summary 1 was genera...
13		TRUE	7_gold	"But it didn't take me lon...	Georgia continues to reg...	San Francisco (CNN) -- ...	A way forward for pregn...	Summary 1	Summary 2 was genera...
8		TRUE	8_gold	The Sahara desert cove...	Editor's note: Christiane...	Editor's note: Christiane...	Amanpour to girls: It's ti...	Summary 2	Summary 1 was genera...
10		TRUE	9_gold	(CNN) -- An American la...	Got a question about eti...	Editor's note: Editor's no...	An open letter to texting...	Summary 2	Summary 1 was genera...
11		TRUE	10_gold	"What happened there i...	U.S. airlines collect $6 b...	(CNN) -- What makes ai...	And the most satisfying ...	Summary 2	Summary 1 was genera...
9		TRUE	11_gold	Warhol relied on a copy ...	The findings are publish...	Prof Dan Goldman expla...	Ant studies to aid design...	Summary 2	Summary 1 was genera...
11		TRUE	12_gold	Google, Amazon and St...	UK fruit growers could e...	UK fruit growers could e...	Apples and pears shapi...	Summary 2	Summary 1 was genera...
16		TRUE	13_gold	The dictator is dead is th...	A landlocked country wit...	Former Argentine militar...	Argentina ex-military lea...	Summary 1	Summary 2 was genera...
6		TRUE	14_gold	GMT / 10:30 a.m. CET a...	Campus life needs to be...	Watch "Iraq: 10 Years O...	Arwa Damon's Iraq: Suff...	Summary 1	Summary 2 was genera...
16		TRUE	15_gold	"It started hopping aroun...	By the beginning of the ...	(CNN) -- An Australian p...	Australian politician injur...	Summary 1	Summary 2 was genera...
10		TRUE	16_gold	On the northern tip of th...	The country has been h...	Bahrain - which name m...	Bahrain profile	Summary 2	Summary 1 was genera...
12		TRUE	17_gold	One ascending balloon ...	For the first time, Mr Ob...	Canadian tourist E Way...	Balloon crash kills touris...	Summary 1	Summary 2 was genera...
9		TRUE	18_gold	Opposition figures are s...	For such a small country...	For such a small country...	Belgium profile	Summary 2	Summary 1 was genera...
9		TRUE	19_gold	(CNN) -- Guess it's a cul...	Ira Foreman was named...	(CNN) -- Guess it's a cul...	Berlin Barbie bummer; f...	Summary 1	Summary 2 was genera...
12		TRUE	20_gold	It said it would treat the...	It is now an independen...	Bosnia-Hercegovina is r...	Bosnia-Hercegovina pro...	Summary 2	Summary 1 was genera...
13		TRUE	21_gold	Country profiles compile...	The average list price fo...	(CNN) -- This was no or...	Can the 'game-changer'...	Summary 2	Summary 1 was genera...
12		TRUE	22_gold	How prepared are we as...	His answer made me la...	Editor's note: John D. S...	Cartoons that scare Syri...	Summary 2	Summary 1 was genera...
8		TRUE	23_gold	The origins of the strugg...	They will not be able to ...	China is one of a numbe...	China joins Arctic Counc...	Summary 2	Summary 1 was genera...

Page 1 of 25 | Displaying rows 1 - 25 of 616

Figure 5.1 Screenshot of the Data section to indicate the Test Question

We randomly set the paired summary as Summary 1 and Summary 2. For example, the first document may have EXP as Summary 1 and LOCAL as Summary 2, and we would have a different placing for the next document. This interchangeable placing made sure that the participants would not be biased towards the same summary name and settings.

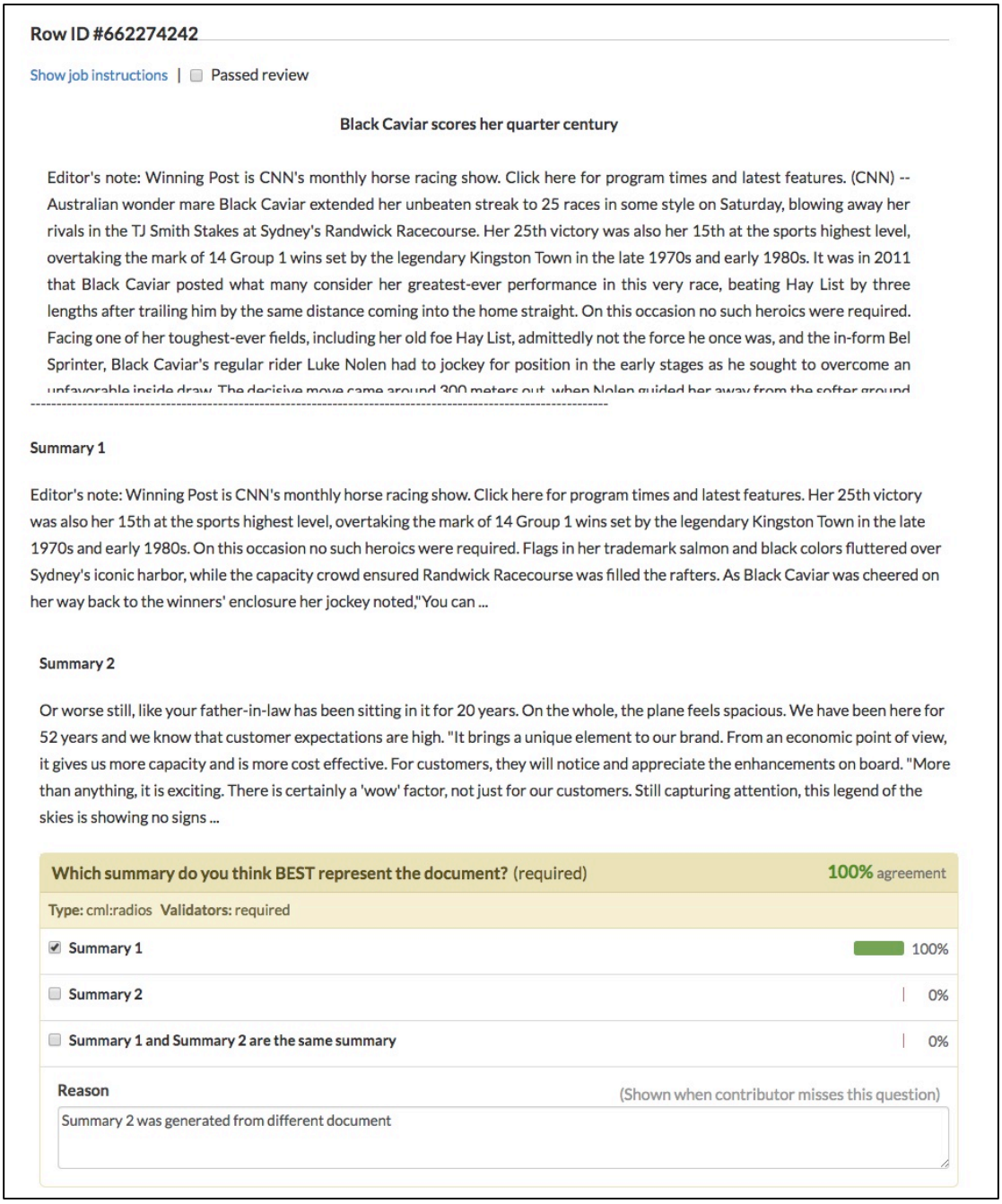


Figure 5.2: Example of test question creation page

Figure 5.2 shows the Test Question creation page with the answers and reasons. And in the Test Question settings, we set two quality controls for all jobs:

1. Each participant must maintain a minimum of 70% accuracy for the test questions throughout the job (to be accepted as 'trusted judgment'), and
2. Each participant must spend a minimum of 45 seconds per page.

If any of these conditions were failed to be followed, the participants would be removed from the job.

5.2.3 Pilot Test

Prior to our main experiments, we ran pilot tests to ensure our instructions were clear and to make sure we gathered the information we wanted. In our first two pilot tests, we provided the participants with two options: Summary 1 or Summary 2 (Figure 5.3), as designed by (Mackie et al., 2014; Yulianti et al., 2015).

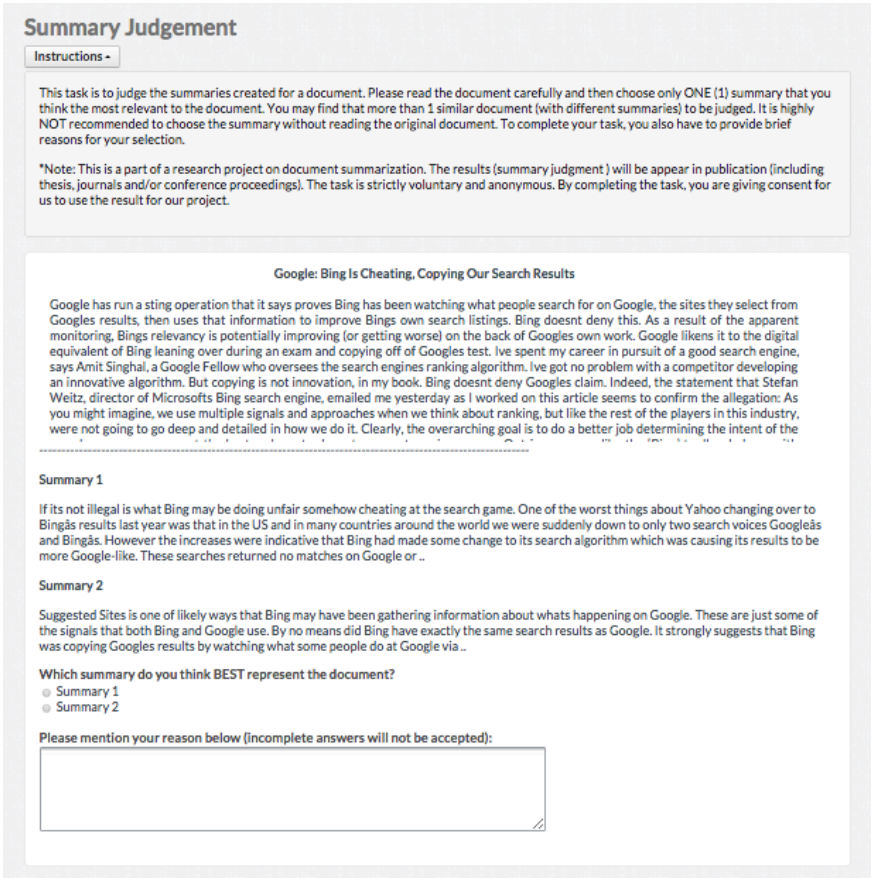


Figure 5.3: The first test pilot screenshot

When we analysed the comments, we noted that there were comments that informed us that the provided summaries are the same. However, since there were only two options (Summary 1 or Summary 2), participants randomly chose answers.

It has been suggested that an option such as "*I don't know*" be included to avoid the participants guessing the answer (Alonso, 2012). Work by Glaser & Schütze (2012) included a third option ('*Neither sentence has a convincing reason'*). We improved our next pilot test by providing a third choice '*Summary 1 and Summary 2 are the same summary'*, as in Figure 5.4.

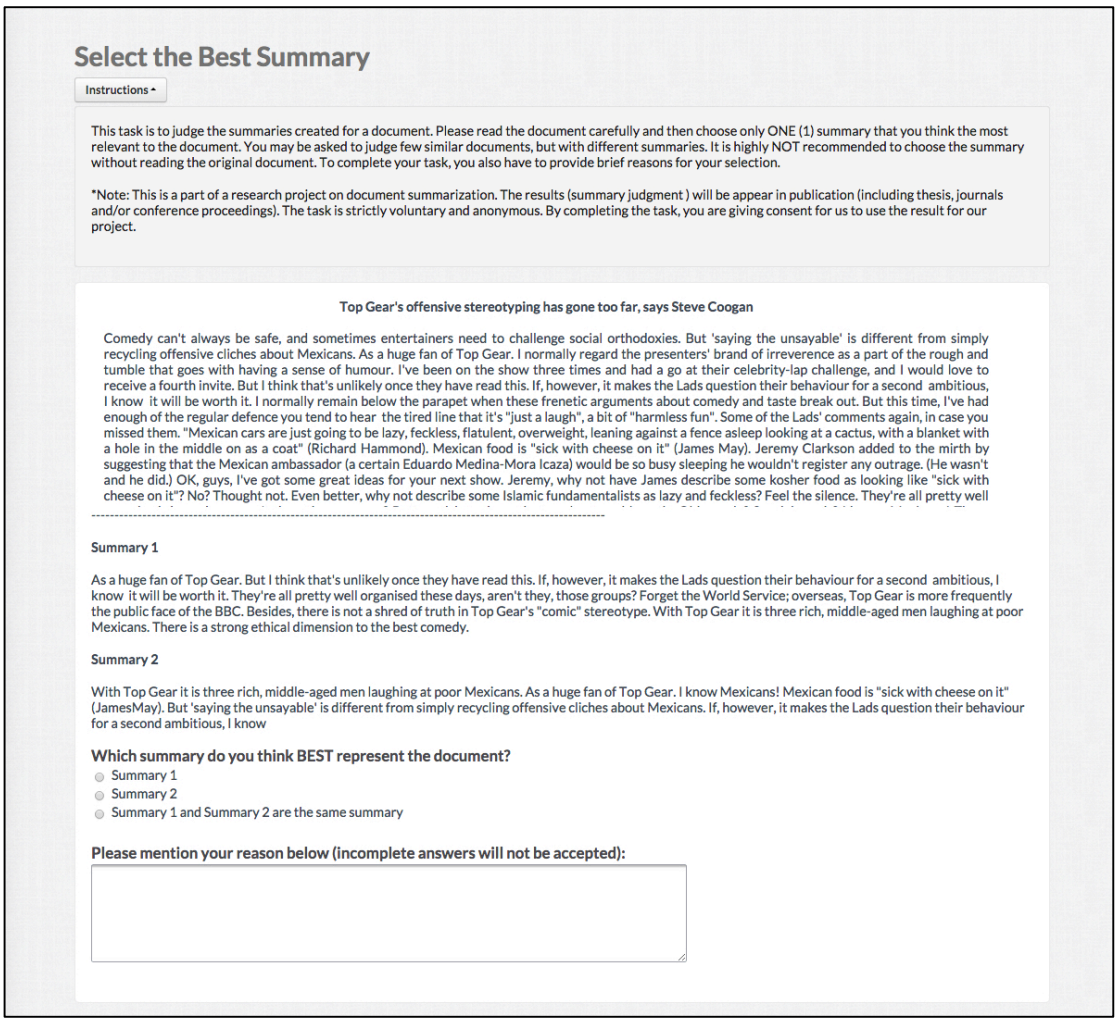


Figure 5.4: Screenshot of CrowdFlower Task

We also discovered participants gave comments in unidentified languages. We were able to identify the background of the participants (from countries such as Vietnam, China, Bangladesh etc.). We excluded these countries in the other pilot tests and in our main experiments. However, note that we did not restrict our experiments to English speaking countries only because we felt it is important to have multi-lingual participants involved in our experiments. This would help us to gather the understanding of the generated summaries from different background.

At the end of our pilot test experiments, we were able to gather the results that we expected; hence we continued our experiments with the dataset as discussed in the next section.

5.2.4 CrowdFlower Setup

In the CrowdFlower job, the participants were asked to judge a set of summaries. We showed them a set of documents each with two summaries created from the document. They were asked to select the best summary and to write why they made their judgement.

Each page shown to participants contained five documents and its paired summaries (5 rows per page). For each document, we required a minimum of 5 participants to judge the summary pair. We paid 15-20 cents per page, where we paid in average AUD260 for each job. We also set a limit of 500 judgements per participant. On average each participant completed 280-300 judgments. Overall, a 92% accuracy was achieved for the test questions. We collected 3,500–6,000 judgments for each job and we were able to get 100% answers within 2-3 days.

5.3 Results

We discuss the participant preferences (as human judgement) and analyze their comments.

5.3.1 Human Judgement for Paired Summaries

Table 5.1 shows the results of the summary judgement for all pairs. We applied a Chi-Square test to look for significances of human judgement between the two settings. For all of the paired settings, we obtained $p<0.0001$.

Table 5.1: Summary Judgement for All Settings (%)

EXP	**LOCAL**	**Same Summary**
55.6	33.4	11.0
EXP	**EXP_s**	**Same Summary**
45.4	29.2	25.4
EXP	**T10.EXP**	**Same Summary**
20.7	23.1	56.2
EXP	**LD.EXP**	**Same Summary**
57.2	29.6	13.2
EXP	**TBS_{sn}**	**Same Summary**
18.86	73.43	7.71
EXP_s	**LOCAL**	**Same Summary**
48.1	40.7	11.2
LD.EXP	**LOCAL**	**Same Summary**
43.2	39.1	17.7

The summaries generated by EXP were chosen as the 'best' compared to summaries generated by LOCAL, EXP_s, and LD.EXP. The LOCAL summaries showed the lowest number as a better summary in all paired settings. For the pair, EXP-T10.EXP, 'Same Summary' was the most chosen option (56%), and T10.EXP was chosen slightly higher than EXP. It would appear that the top 10 tweets result in the same sentences being selected as those selected by EXP.

Comparing EXP-TBS_{sn}, TBS_{sn} was notably chosen as a better summary (74%), as was shown in the Recall ROUGE scores of chapter 4. Consequently, we examined participants' comments EXP-TBS_{sn}, to understand the reasons for the strong preference as a good summary.

5.3.2 The Condorcet Method

We next produced a ranking from the paired evaluation. We chose the Condorcet ranking method (Baker, 1975), where a winner of a pair-wise evaluation is determined by calculating the majority rule of the pairing. The Condorcet method had been used to determine winners in an election, and also applied for ranking in IR (Volkovs, Larochelle, & Zemel, 2012; Volkovs & Zemel, 2014; Wei, Gao, El-Ganainy, Magdy, & Wong, 2014) and in document summarization,(Palshikar, Deshpande, & Athiappan, 2012). We also found similar work in Mackie et al. (2014), where they reported on ranking the preference of microblog summaries using Condorcet.

We removed the T10.EXP and TBS_{sn} settings because both settings are only paired once. We also removed "The Same Summary" result because we were only interested in the selected best summary. We recalculated the percentage of judgements and used this for our ranking method calculation.

The recalculated results are as follows:

Recalculated judgement	Winning settings	Winning Vote
EXP (62) and LOCAL (38)	EXP > LOCAL	62
EXP (61) and EXP_S (39)	EXP > EXP_S	61
EXP (66) and LD.EXP (34)	EXP > LD.EXP	66
EXP_S (54) and LOCAL (46)	EXP_S > LOCAL	54
LD.EXP (52) and LOCAL (48)	LD.EXP > LOCAL	52

Based from these results, we identify a 'winning' summary setting; wherein the Condorcet voting system, a 'win' occurs when a candidate is preferred by a majority of voters. Based from the percentage preferences above, we identified the win-lose pair to generate a voting matrix.

The count for all possible "votes" is shown in Table 5.2, where each row represents the winner of a preference and each column represents the loser. Each cell represents the results of the pairwise comparison, which is the total number of winner 'wins' from all other comparisons. In Condorcet voting system, only the winning vote is used to calculate the Condorcet winner. For example, EXP wins all three paired comparisons, so in all cells of EXP's row, the total vote was 189 (62+61+66).

For LOCAL, the calculation is more complex. We can see that LOCAL loses to both EXP_s-LOCAL and LD.EXP-LOCAL, so all of EXP_s and LD.EXP winning votes (54+52) are given to EXP. For the third column, again all of LOCAL's opponent winning vote is used to calculate the total number of vote for LOCAL-LD.EXP paired match, that is EXP-LOCAL (62) and EXP_s-LOCAL (54). The same applies to all columns and the results are in Table 5.2[3].

Table 5.2: Input Table for Condorcet Matrix

Option	EXP	LOCAL	LD.EXP	EXP_s
EXP	-	189	189	189
LOCAL	106	-	116	114
LD.EXP	52	118	-	118
EXP_s	54	115	115	-

Based from the input table (Table 5.2), it is clear that EXP is considered as the 'Condorcet Winner' as EXP beats all of its opponents.

In order to rank the summary settings, we applied the Ranked Pairs method (Tideman, 1987) to the Condorcet voting results. The results are shown in Table 5.3, where we only considered the winning (higher number of judgement). The reason for this is we want to create the '*defeat*'[4] rules (as illustrated in Figure 5.5).

[3] To calculate the Condorcet method, we used the tool provided in http://condorcet.ericgorr.net/

[4] We defined '*defeat*' as a method outperformed another method.

In Table 5.3, only the winning votes for each pair (from Table 5.2) remains on the table. For example, EXP wins all of its paired votes, so all votes in EXP's row (the winning vote) remains. As for LOCAL, it loses to all its component, thus all votes in its row are changed to 0. LD.EXP lost to EXP (vote change to 0) but still wins with LOCAL and EXP_s. EXP_s only outperformed the LOCAL summaries.

Table 5.3: The Defeat Matrix

Option	EXP	LOCAL	LD.EXP	EXP_s
EXP	-	189	189	189
LOCAL	0	-	0	0
LD.EXP	0	118	-	118
EXP_s	0	115	0	-

The results in Table 5.3 is used to create the defeat rules in Figure 5.5:

The Defeat rules

- EXP outperforms EXP_S
- EXP outperforms LOCAL
- EXP outperforms LD.EXP
- LD.EXP outperforms LOCAL
- LD.EXP outperforms EXP_S
- EXP_S outperforms LOCAL

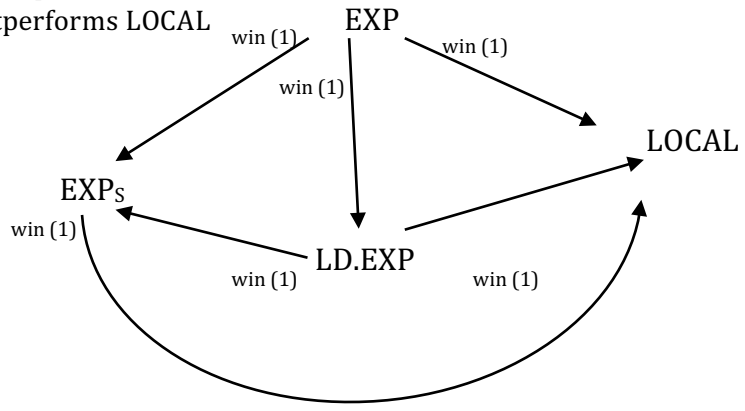


Figure 5.5: Pairing winning votes for AG settings. The arrows that point away show the winning path.

We can decide the winning setting based on the winning votes, as in Table 5.4:

Table 5.4: The Ranking based on the winning votes

Option	Total Win	Rank
EXP	3	1st
LOCAL	0	4th
LD.EXP	2	2nd
EXP_s	1	3rd

The settings were ranked EXP, LD.EXP, EXP_s and LOCAL. This result showed that the summaries generated with the support by its tweets are chosen as a better summary compared to LOCAL summaries. The EXP setting received the most vote as the best summary.

5.3.3 Condorcet Ranking for Summary Examples

In Chapter 4, we manually analysed the summaries created by different Affinity Graph approaches, where we randomly selected four documents that represent different lengths and the number of related tweets. Table 5.5 shows the characteristics of the chosen documents. In the analysis, we found that DocID 2 and 311 had better summaries compared to the shorter documents. However, for DocID 311, the summary content was different for each Affinity Graph setting.

For the shorter documents DocID 55 and 426, the summaries contained many words found in the tweets. However, since the documents are much shorter (<10 sentences), most of the tweets contained the main topic of the documents, thus generated a more topic-focused summary.

We examined the Condorcet voting system for four documents (see Table 5.1). Table 5.5 shows that the tweet-biased Affinity Graph approaches showed a ‘win’ in Condorcet voting system when used for DocID 2 and DocID 311.

Table 5.5: Human Judgement Ranking for Summary Evaluation

Documents	**# sentences**	**# tweets**	**Higher ranked Affinity Graph Settings (based on Condorcet Voting system)**
DocID 2	208	80	EXP
DocID 311	694	11	EXP_S
DocID 55	6	21	LOCAL
DocID 426	9	10	LOCAL

For DocID 2, the EXP setting was chosen as the winning summary compared to the others (EXP_S, T10.EXP, LD.EXP, and LOCAL). We can see that in Chapter 4 (Figure 4.12) the EXP and EXP_S settings generated the same summary, and in the CrowdFlower paired test, the participants judged that 'Summary 1 and Summary 2 are the same summary' with 5 votes. But based on the Condorcet voting system, EXP is chosen as the winner because EXP wins in all of its pairwise comparisons. For DocID 311 (Figure 4.13), the EXP_S summary received a high ranking in the Condorcet voting system. Our analysis suggests this is because the summary contained one sentence on Irish banks and Irish property, where each topic appeared in the reference summaries. This suggests that for long documents, more tweets help to improve the summaries regardless of the expanded document settings in the Affinity Graph algorithm.

For shorter documents (DocID 55 and DocID 426), the LOCAL Affinity Graph approach generated a preferred summary. This suggests that for shorter documents less external information is needed to select appropriate sentences.

Based on the manual analysis and the human judgement (Condorcet voting), we discovered that there are two conditions where the tweet-biased Affinity Graph settings could improve summaries: (1) the length of the document (the longer the better) and (2) the number of related tweets (higher number of tweets are better).

5.3.4 Analysis on Participant Comments

We asked participants to explain the reason for their chosen summary and evaluated 19,186 comments across all judged summary pairs. We have seen only a few past works that sought and discussed participants comments (Kushniruk et al., 2002; Mackie et al., 2014; Yulianti et al., 2015). Other works that have discussed participant comments can be found in Sanderson et al. (2010), where they discussed on search engine ranking; Kim, Oh, & Oh (2008) and Maglaughlin & Sonnenwald, (2002) on participant relevance criteria, Barry (1994) on evaluating information in a document and Savolainen & Kari (2013) on web-searching.

We took a qualitative approach to the free text comments adapting an existing inductive analysis for qualitative data method (Thomas, 2006):

1. Preparation of the raw data.

   Firstly, we clean the data, where we identify if there are comments just containing symbols (!@#$%^&*) and assigned the comments as Spam.

   In this stage, we also noticed that there are comments that are identical (e.g.: "*Summary 1 is more relevant"*, "*This explains more"*, etc.). We grouped such comments and counted the number of times they appeared.

2. Close reading of the text.

   Next, we close read the comments and identified keywords (e.g.: *similar*, *relevant*, *same*, *like*, *better*, *detail*, *represent*, *important*, etc.). We grouped comments by an identified keyword. We repeated this step for a few times to make sure we can be consistent with our groups and themes.

3. Creation of categories.

   Once we have identified themes based on the keyword of the comments, we identified larger categories. We merged keyword groups into categories. For example, comments from groups

with keywords like "*More key points/main points*", were grouped together with "*Relevant*" and "*More accurate*" and categorized as groups that discussed the *Content* of the summary.

4. Revision and refinement of the category.

We searched for the topic of the comments and compared our categories with other work. We also made multiple readings and interpretations of comments until we were able to categorize all comments.

Note that when assigning a comment only belong to one category. For the few verbose comments that contained more than one category keyword (less than 5% of the total comments), we assigned the comment based on the more specific keyword. For example, the comment "*summary 1 does it better with less words"*, will be assigned to the "*less word*" category, rather than the "*better*" category.

5.3.4.1 Comments Category

We categorized the comments/reasons into seven main groups: ***Topic Discussed***, ***Document-Summary Similarity***, ***Presentation***, ***Same Summary, Preference***, ***Not Classifiable,*** and ***Spam***.

Presentation and Topic Discussed were frequently used in participant behaviour analysis for preferring one document ranking over another (Kushniruk et al., 2002; Sanderson et al., 2010) or in a participant-defined relevance criteria for web-searching (Barry, 1994; Kim et al., 2008; Maglaughlin & Sonnenwald, 2002; Savolainen & Kari, 2013). In microblog document summarization, Mackie et al. (2014) identified five main categories for participant comments (Informative, Readability, Length, Sentiment, and Tweet-Specific); whereas Yulianti et al. (2015) identified 28 categories which were later combined into three categories (Content, Writing/Presentation, and Flow). Summary content and presentation are commonly identified as important features as discussed in previous work. However,

we split the comments on content into two different groups: *Topic Discussed* and *Document-Summary Similarity*.

In **Topic Discussed**, we categorized the comments that mentioned if the topic of the local (original) document appeared in the summary. In this category, we see that the participants can either identify the main topic/points of the local document (*On Topic*) or they found relevant or important information in the summary (*Relevant Information*). We found that this category has the same definition with the *Informative* and *Sentiment* from Mackie et al. (2014) and *Content* from Yulianti et al. (2015).

In the *Relevant Information* sub-category, written comments contained the words "*relevant*", "*more detail*" and "*more information*". We assumed that the participants understand the content of the local (original) document, thus, they agreed that the summaries contained the specific information of the local document.

For comments such as "*It's about Obama*", "*on topic*" or "*about twitter*" we assume participants are stating that the summary contained the topic of the local document. We also included negative comments about the contents in this category. Here, the participants commented on what they did not like in one of the paired summaries (e.g.: "*off topic*", "*different issue*").

Another category that we identified from the participant comments is **Document-Summary Similarity**. This category is different from **Topic Discussed** because the focus of the participants is on the words of the summaries rather than the topics. de Oliveira (2005) discussed that the quality of a summary is related to its similarity with the original document. We note that this category has not been mentioned in other similar work, but we believed that this is an important feature for document summarization.

For this category, most of the comments state "*Summary_1 text is more similar to original text then Summary_2"* or "*Summary 2 is the same with the text*". Other comments include "*Summary 1 better represents the document*" or "*better represent*". Negative comments such as "*Summary_1 did not represent the document*" or "*1 is different from text*" are also included in this category.

In the **Presentation** category, we put comments that describe the readability of the summary. Participants commented on the writing, structure, and order of the sentences in the summary, the same definition in work by Mackie et al. (2014) and Yulianti et al. (2015). Though note, Mackie et al viewed *Readability* and *Length* as different features and in the later work, defined *Writing/Presentation* and *Flow* as two different categories.

Participants commented on the length of the summary (e.g., "*this is short*", "*Summary 2 is longer than Summary 1*"), writing (e.g., *"Summary 1 is more understandable."*), summary structure (*"Summary 1 is well-structured"*, *"The second summary has a better flow to it."*) and order of the sentences (*"Summary 1 has sentences in correct manner and good for judgement"*, *"In the right order")* in the summary.

The **Same Summary** is defined as the summary extracted exactly the same sentences from different paired documents. This included comments such as "*both are same*", "*they are the same summary*" or "*Exactly the same*".

We found a small number of participants that commented a **Preference** without reason: "*I choose summary 1*" or "*This is the best summary for me*" or a single word like "*like*".

For comments such as "*this is better*", "*good summary*", "*OK*" or "*Correct*", we created the **Not Classifiable** category.

In **Spam**, we identified comments that are irrelevant and do not reflect any reasons for the participant to choose their summary. We also checked about 2000++ comments and identified that they were copied sentences from the local documents. Therefore, we categorized these comments as uninformative. Table 5.6 shows examples of the comments from the CrowdFlower.

As Table 5.6 shows, there is a difference between the two categories where the participants describe relevance or similarity to the main documents: relevance is more a reference to the meaning of the text, whereas similarity is more focused on word overlap between summary and document.

Table 5.6: Examples of the Comments and Category

Category	Reasons	Comments Example
Presentation	Summary length	*"summary 1 does it better with less words"* *"Short and Informative"* *"it's short but it's better than the other one."*
	Quality of the presentation (impressive/attractive/easy to understand)	*"Summary 2 ends with a complete sentence."* *"They are just the same, just different presentation."* *"Summary 1 is well-structured"* *"The second summary has a better flow to it."* *"Summary 1 has sentences in correct manner and good for judgement."*
	Negative comments ("Not well written"/"weird")	*"Summary 2 starts off from the middle of the article so it doesn't make sense"* *"Summary_1 starts from the end of article!"* *"Summary_1 is just a collection of random sentences and weird."* *"1st text isn't a summary, it's just random words"*
Topic discussed	On topic	*"summary 2 describes how twitter may reach 150$ million in advertisements better"* *"Mentions Facebook which is important to the article"* *"Better at mentioning the details of the IPO"* *"It's about Kate Spade and the thumblr"* *"although the second one mentions part of article, only the first one summarizes apples new APP"* *"I was able, through this summary, more easily understand the topic"*
	Relevant information included in the summary	*"Because summary 1 is more relevant."* *"More detailed information".* *"summary 1 has content and official statements."* *"summary 2 managed to explain 2 of the 22 stories promised by the main article"*
	Negative comments ("insignificant points / weird")	*"summary 2 is off topic"* *"summary two is from a very different issue"* *"Summary_2 is not to the point."* *"Summary 1 doesn't mention the basic plan and may be considered misleading on price."*
Document-Summary Similarity	Similarity with the document	*"Summary_1 text is more similar to original text then Summary_2."* *"This is best related to the text."* *"because summary 2 is quite similar to above given summaries"* *"Summary 2 is more suitable because it's more represent the content of document."*
	Negative comments ("Different from text"/"Not representing")	*"summary 2 is not related"* *"summary 1 is not from the document"* *"summary 2 did not mention the title of the book."*
Same Summary	Both summaries identified as	*"Summary 1 and Summary 2 are the same summary"*

	the same.	*"Both are same summary."* *"Exactly the same, word for word."*
Not classifiable	Ambiguous reason ("is better", "clear")	*"Summary is better."* *"Is much clear"* *"More okay"*
Preference	Preference comments	*"I choose summary 2."* *"I prefer 1"* *"This is the best summary for me"* *"Because the answer I choose it's correct..."*
Spam	Spam comments	*"Happy exploring!"* *"No comments...."* *"….."*
	Copied text from local document	*"Right now, a video featuring a Brazilian taxi driver doing a spot-on Michael Jackson impression is going viral, and it's likely only a matter of time before the job offers start rolling in."* *"Shaxson shows how the world's tax havens have not, as the OECD claims, been eliminated, but legitimised;"*

5.3.4.2 Comments-Judgement Analysis

In Figure 5.6, ignoring Not Classifiable and Same Summary categories, the main reasons for the participant to express preferences are Topic Discussed and Document-Summary Similarity. In EXP_s-LOCAL, we can see that the two categories have a similar percentage (29.2% vs. 29.7%). EXP_s-LOCAL also showed the highest percentage for Presentation (13%) compared to the other pairs. Based on the human judgements (Table 5.1) and comments, the summaries generated by EXP_s and LOCAL only appear to have small differences between them. Thus, we assumed that the summaries generated only with its sentences-tweet similarity score are equally good with LOCAL summaries.

The pair EXP-T10.EXP showed the highest percentage of The Same Summary (45.5%). This also agrees with the preference results in Table 5.1. The additional tweets available to EXP did not appear to make much of a difference. This result agrees with the results in Table 4.5 in Chapter 4, where the recall and precision score showed a small difference between EXP and T10.EXP.

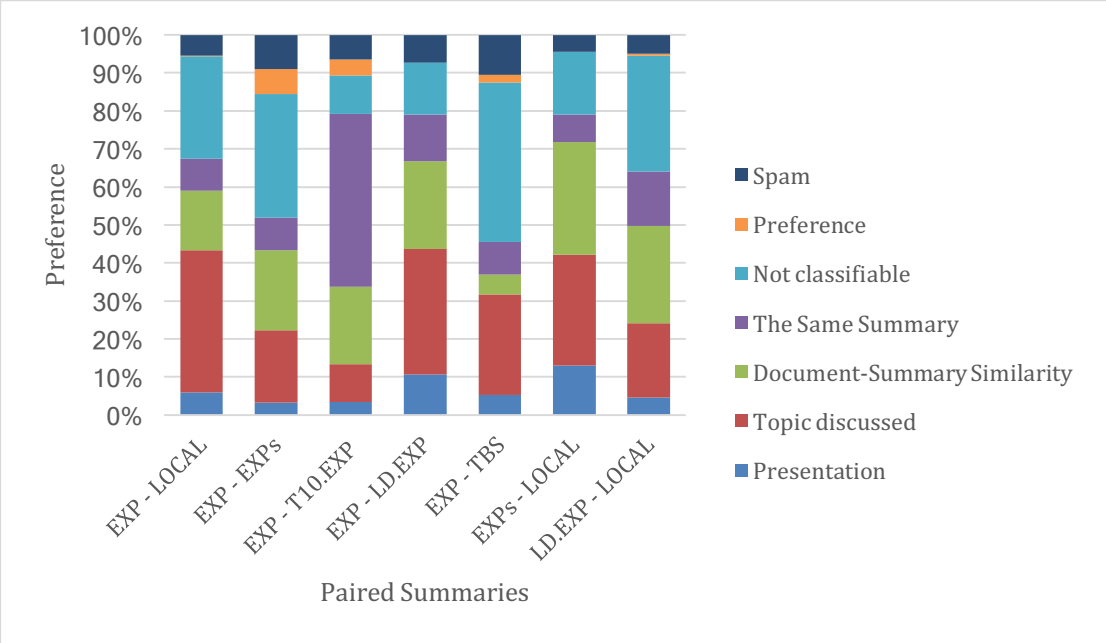


Figure 5.6: Participant Comments Category for All Settings (%)

Table 5.7 shows a more detailed result of the participant comments analysis. In EXP-LD.EXP, 29.47% of the participants mentioned that relevant information appeared in the summaries generated by EXP. Whereas, in EXP-LOCAL, the participants identified that the chosen summary has included the main document topic (On topic – 11.27%), the highest percentage for On Topic. This result shows that EXP appears able to identify the topic discussed in the local document and extract more relevant sentences, as agreed by the participants when compared with the LOCAL and LD.EXP summaries.

For the comparison between EXP and TBS_{sn}, Topic Discussed showed a higher percentage of the participant comments. In Table 5.7, 24.66% of the participants stated that summaries by TBS_{sn} contained more relevant information from the original document and is able to extract better sentences compared to EXP. Again, this result agreed with the results in Table 4.5 (Chapter 4), where we showed TBS_{sn} gave better Recall score compared to EXP.

Table 5.7: Participant Comments' Category

		EXP - LOCAL	EXP - EXP_S	EXP - T10.EXP	EXP –– LD.EXP	EXP - TBS_{sn}	EXP_S - LOCAL	LD.EXP - LOCAL
Presentation	Summary length	0.23	0.12	0	0.04	0.28	0	0
	Quality of the presentation	5.16	2.93	3.39	10.65	4.66	12.97	4.32
	Negative comments	0.50	0.19	0.08	0	0.34	0	0.21
Topic discussed	Relevant information	25.36	15.03	6.19	29.47	24.66	25.59	12.75
	On topic	11.27	3.51	3.46	3.50	1.69	3.50	6.91
	Negative comments	0.84	0.42	0.27	0.04	0.07	0.11	0.04
Document-Summary Similarity	Similarity with document	14.82	21.05	20.12	22.95	5.28	29.66	25.26
	Negative comments	0.80	0	0.16	0.11	0	0	0.35
The same summary	The same summary	8.48	8.71	45.53	12.30	8.66	7.14	14.17
Not classifiable	Ambiguous reason	26.93	32.42	10.04	13.59	41.98	16.61	30.64
Preference	Preference	0.19	6.63	4.28	0	1.86	0	0.43
Spam	Spam	5.42	8.98	6.50	7.36	10.52	4.42	4.92

We collated the results from Figure 5.6 to identify the preference reasons for each of the summary setting (Figure 5.7). We included Not Classifiable in the analysis to see how many of the participants did not state their specific reason for their chosen summaries.

In Figure 5.7, we can see that the main reasons for the participants to choose the summaries generated by tweet-biased approach are the Document-Summary Similarity and the Topic Discussed. Both EXP and EXP_S were selected by the participants because it included relevant topics in its summary. However, EXP_S was also selected because the summaries generated by EXP_S are almost similar to the original (local) document. The EXP_S summaries also showed a balanced result between Document-Summary Similarity and Topic Discussed compared to EXP.

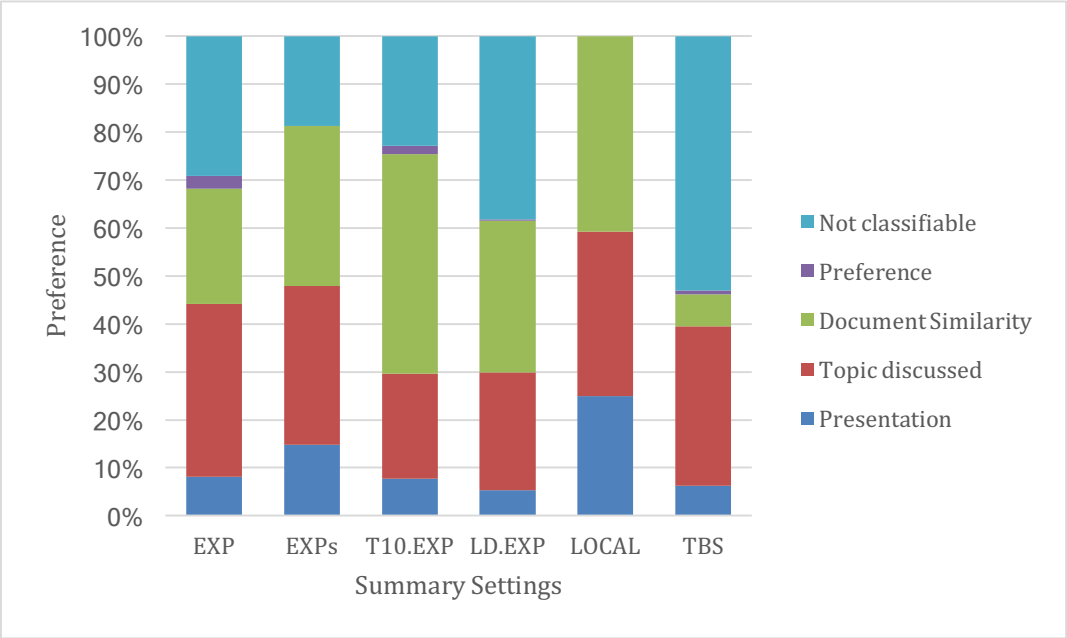


Figure 5.7: Comments Category for Different Summary Settings (%)

The summaries generated by T10.EXP showed a high preference (42.4%) of Document-Summary Similarity. Here we can see the difference between EXP and T10.EXP: a greater emphasis on Topic Discussed for EXP and Document Similarity for T10.EXP. The LOCAL summaries have the highest preference due to Presentation.

For TBS_{sn}, Not Classifiable is the highest reason for the participant to choose its summary. Most the participants agreed that Topics Discussed is the best reason for them to choose TBS_{sn}. We also applied a Chi-Square test to look for significance between all categories of the participant comments. Assuming an even distribution between the categories, for the statistical test, we obtained $p<0.0001$ for all categories in each setting.

5.4 Baseline vs. Affinity Graph's Tweet-Biased Summaries

Our findings in Table 4.5 (Chapter 4) showed that the basic baseline summaries (the first 100 words from the document – BASELINE 1) performed significantly better than all of the Affinity Graph/Tweet Biased Summarization in terms of its recall and precision scores. We compared human judgements between Baseline and Tweet-Biased Summaries using the same experiment settings as discussed in sections 5.2 and 5.3. To ensure consistency in the human judgement experiments, we re-ran one of the paired settings discussed in 5.3 (the EXP-LOCAL settings). The latter experiment showed that the human judgement results were not much different from the results reported earlier[5].

Table 5.8 shows that BASELINE summaries were selected as the 'best' summary by participants, except when paired with EXP. EXP wins 50.5% of the preference compared to BASELINE summaries (30.4%), and only 19.1% of the summaries were identified as "Same Summary'.

Table 5.8: Human Judgement for Summaries (Baseline vs. different Affinity Graph's setting (%))

EXP	**BASELINE**	**Same Summary**
50.5	30.4	19.1
EXP_s	**BASELINE**	**Same Summary**
11.4	60.0	28.6
LD.EXP	**BASELINE**	**Same Summary**
4.7	76.7	18.6
LOCAL	**BASELINE**	**Same Summary**
4.5	83.9	11.6
T10	**BASELINE**	**Same Summary**
15.0	70.6	14.4
TBS	**BASELINE**	**Same Summary**
0.81	64.9	34.3

[5] In Table 5.1, we reported that the EXP was chosen 55.6% over LOCAL (33.4%) and Same Summary (11.0%). From our repeated experiment, the EXP was chosen 55.4%, LOCAL (31.2%) and Same Summary (13.4%), a discrepancy of ~ 0.2% to 2%.

In order to rank the summaries (to see if EXP still wins when we include BASELINE summaries in the summary judgement comparison), we applied the Ranked Pairs method, as discussed in 5.3.2. For this, again we recalculated the human judgement results by eliminating the vote for 'The Same Summary'. This time we included T10.EXP and TBS_{sn} to rank all summaries together with the BASELINE summaries.

Table 5.9: Condorcet Matrix for Baseline and all AG settings

Option	BASE	EXP	EXP_S	LOCAL	LD.EXP	T10	TBS
BASE	-	356	418	418	418	418	418
EXP	384	-	384	384	384	331	304
EXP_S	115	114	-	175	175	175	175
LOCAL	168	190	198	-	200	252	252
LD.EXP	118	138	194	194	-	194	194
T10	53	124	124	124	124	-	124
TBS	80	145	145	145	145	145	-

Table 5.9 shows the pairwise results based on the preference for experiments in Table 5.1 and 5.8 (for the EXP-LOCAL result, we used the result from the first human judgement experiments as in Table 5.1). Based on the result in Table 5.9, we produced the Defeat Matrix (Table 5.10), where $^{(1)}$ shows the 'win' based on the pairwise number preference of the participants.

Table 5.10: The Defeat Matrix

Option	BASE	EXP	EXP_S	LOCAL	LD.EXP	T10	TBS
BASE	0	0	$418^{(1)}$	$418^{(1)}$	$418^{(1)}$	$418^{(1)}$	$418^{(1)}$
EXP	$384^{(1)}$	0	$384^{(1)}$	$384^{(1)}$	$384^{(1)}$	$331^{(1)}$	$304^{(1)}$
EXP_S	0	0	0	0	0	$175^{(1)}$	$175^{(1)}$
LOCAL	0	0	$198^{(1)}$	0	$200^{(1)}$	$252^{(1)}$	$252^{(1)}$
LD.EXP	0	0	$194^{(1)}$	0	0	$194^{(1)}$	$194^{(1)}$
T10	0	0	0	0	0	0	0
TBS	0	0	0	0	0	$145^{(1)}$	0

The results in Table 5.10 is used to rank the summaries (in Table 5.11) based on their number of preference (or votes) of each paired summaries. We can see that, EXP has the most wins (6) of the pairwise preferences, followed by BASE, LOCAL, LD.EXP, EXP_s, TBS and T10.

Table 5.11: The Ranking based on the winning votes

Option	Total Win	Rank
BASE	5	2
EXP	6	1
EXP_s	2	5
LOCAL	4	3
LD.EXP	3	4
T10	0	7
TBS	1	6

Only EXP and BASELINE win nearly all of its pairings: EXP loses to TBS and BASELINE loses to EXP. However, in the Condorcet voting system, EXP is considered as the winner because it wins in all 6 matches – BASELINE loss one vote, that is the BASELINE – EXP match. The last rank was T10.EXP, because T10.EXP only wins once, when paired with EXP but only by 53; thus the number of votes was not enough for T10.EXP to win in the Defeat Matrix.

For the comments analysis in BASELINE and Affinity Graph summaries, we were only interested in the reasons for preference for BASELINE and EXP. Thus, we removed the "Same Summary" (which is 15.5% of all comments) and "Spam" (11.48% of all comments). Further discussion on the "Same Summary" will be in Section 5.6.

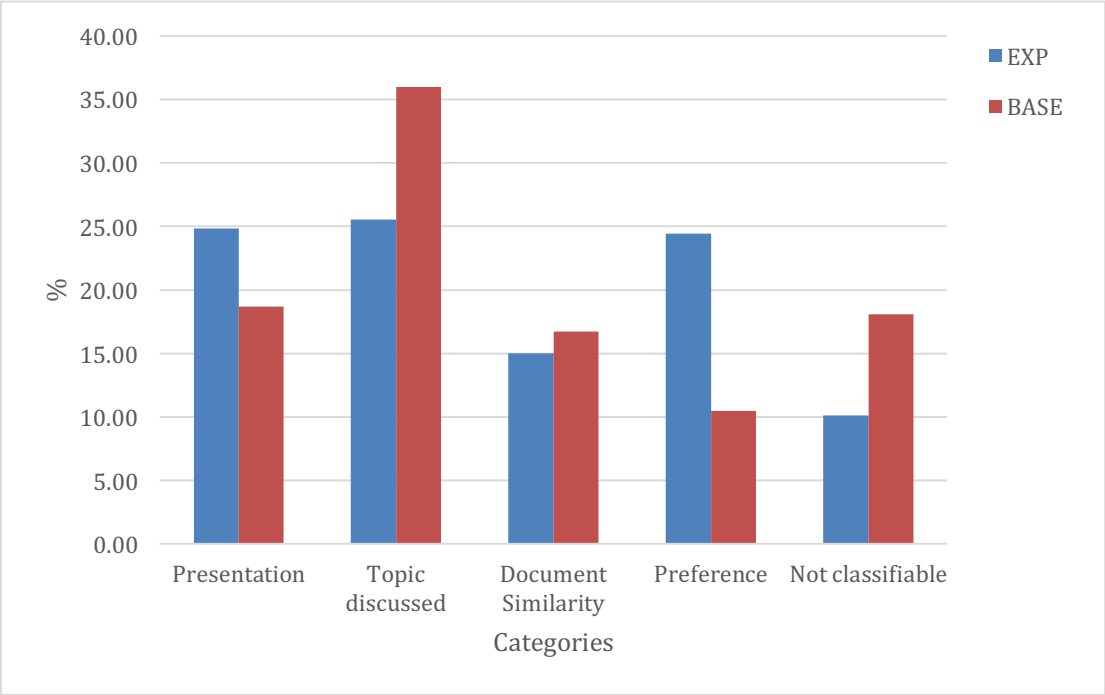


Figure 5.8: Comments analysis for EXP and BASELINE

In the BASELINE summaries, we accumulated the comments from all pairwise summaries; except for EXP-BASELINE, where we used the EXP-BASELINE comments summaries for EXP (because EXP wins the pairwise comparison). From Figure 5.8, we can see that the top reason for the participants to choose BASELINE is the Topic Discussed, where the same topics were identified in both summary and document. BASELINE has slightly more comments (2%) that identified the summaries were similar with the document, and 7% difference in the Not Classifiable category (see Table 5.12).

Table 5.12: Participant's Comments for their elected summaries

		EXP (%)	**BASE (%)**
Presentation	Summary length	0.10	0.29
	Quality of the presentation (impressive/attractive/easy to understand)	24.51	18.06
	Negative comments ("Not well written"/"weird")	0.41	0.45
Topic discussed	Relevant information included in the summary	3.13	17.35
	On topic	22.15	18.05
	Negative comments ("insignificant points / weird")	0.41	0.70
Document Similarity	Similarity with the document	14.87	16.49
	Negative comments ("Different from text"/"Not representing")	0.15	0.43
Not classifiable	Ambiguous reason ("is better", "clear")	10.05	17.82
Preference	Preference	24.21	10.35

Note, even though BASELINE has the most comments in Topic Discussed, participants identified that the summaries generated by EXP were more related to the documents (On Topic – 22.15%). EXP also showed a high percentage in the quality of the presentation, such as "*The second summary has the information more condensed*" and "*The summary 2 is easier to understand than the first summary*"; and also for Preference, such as "*for me Summary 2 will be the best of the above choice*" and "*I prefer Summary 1*". This shows that the summaries generated were easy to understand and more preferable than BASELINE.

5.5 Comments Category Analysis

Based on the comments categories in Table 5.5 and the human judgement results in Table 5.1 and 5.8, we collated results to identify the main reasons for the participants choosing their summary. For this, we added up the total number of comments for three categories only: Topic Discussed, Document-Summary Similarity, and Presentation. This is because we considered the other categories (Same Summary, Preference, Not Classifiable and Spam) would not reflect any features for generating a summary. And then for each category, we calculated the percentage as shown in Figure 5.9.

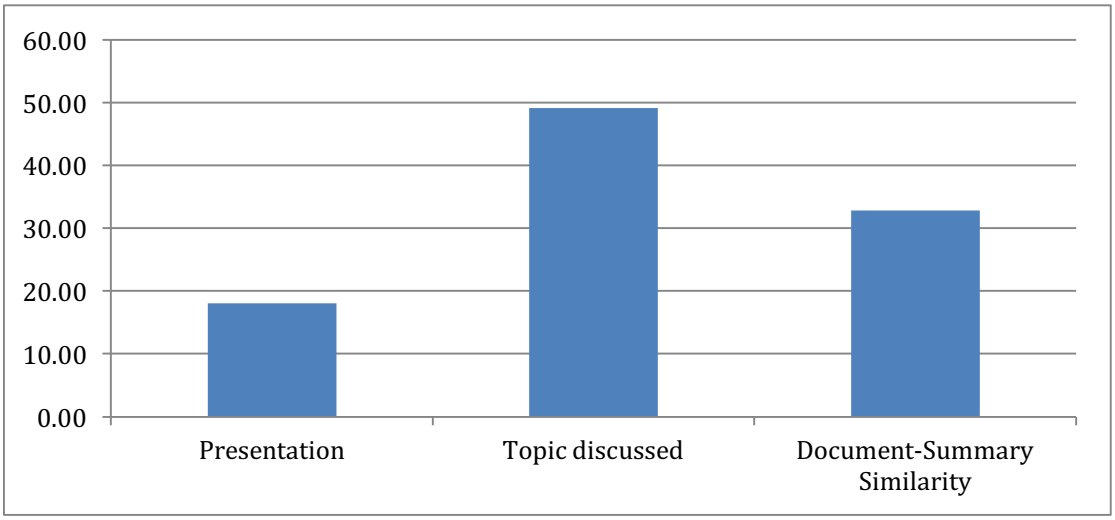


Figure 5.9: Reasons for choosing summaries as The Best by participants

We can see that Topic Discussed is more preferred compared to Document-Summary Similarity and Presentation. We believed that the participants would be able to detect the relevant information in the summaries and this would make an important feature to generate a summary. Thus for a document that has more than one topic in its content, identifying the main topic to be summarized would help to create a summary that is 'participant favourable'. This would be a good feature for a query-biased summarization.

The Document-Summary Similarity is also an important feature. Most automated evaluation metrics (ROUGE, BLEU) measure such similarity. It is apparent from Figure 5.9, however, that the content of the summaries is the most important feature. The finding is consistent with the study by Mackie et al. (2014) and Yulianti et al. (2015) who both examined participants' comments and discovered that the content of the summaries were important preference features.

Presentation of the summaries gave the least preferred reason to choose a summary, however, we believed that the order of the extracted sentence is equally important so that the summary generated are more understandable and thus will be more preferred by the participants.

5.6 Same Summary Analysis

As discussed earlier in the chapter, the main objective of the participant-preference experiments is to see if the participants were able to identify whether the summaries that were generated by various techniques produced the same or similar summaries. This is an important discussion, because we have identified few summaries with the same ROUGE scores, but have different content, as discussed in Chapter 4. We hypothesized that this is caused by *granularity* and *semantic equivalence;* the problem discussed by Nenkova & McKeown (2011), where they identified both problems as the disadvantages of automated evaluation in summarization problems.

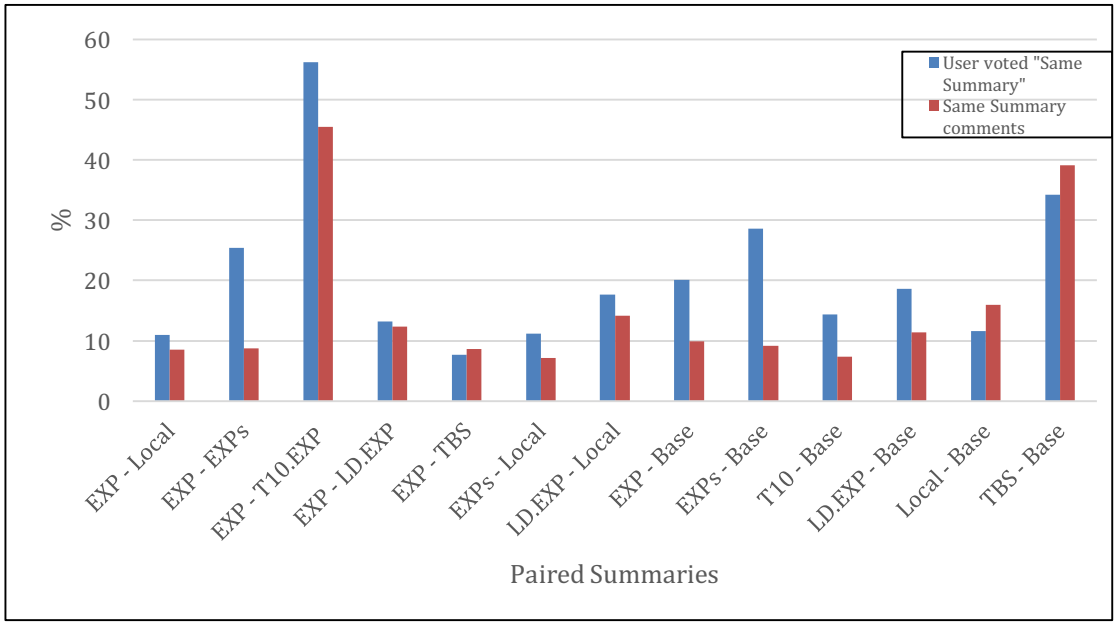


Figure 5.10: The results on the participants who voted "Summary 1 and Summary 2 are the same" and participants' comments that the summary is the same.

In Figure 5.10, we can see similar patterns for the voting results and participants' comments. The pair EXP-T10.EXP and TBS-BASE which has a high vote in "The Same Summary" also has the same high percentage of comments that the summaries are the same.

However, there is an exception in the pair of EXP-Base and EXP_s-Base; both have a high 'Same' option votes, but their 'Same' comments were less than 10% of the overall comments. We further analysed the comments and found out that for both paired settings, the 'Ambiguous' and Preference comments were 38% and 40% respectively. We assumed that the participants were not interested to further commented on their choice because the reason was obvious (it is the same summary).

This also occurs in the EXP-EXP_s, where we can see that the difference between the 'Same Summary' vote and the comments are high (more than 15% differences). Again, further analysis of the comments also showed the same reason as above, the 'Ambiguous', Preference and Spam comments were high - 32.5% (see Table 5.7).

This discussion showed that the "*Same Summary*" option given to the participants gave a valuable insight on the user preference experiments. It shows that the participants were able to identify an exact similarity between the paired summaries. This would help us to further analyse the comments and differentiate the "*Same Summary*" category with the Document Similarity category.

5.7 CrowdFlower Do's And Don't

We believed our experiment for evaluating document summarization using human judgement gave more reliable results compared to the ROUGE evaluation. We also identified features participants used to select their preferred summaries. We find that human judgement via crowdsourcing platform could provide a reliable evaluation metric for document summarization.

In using crowdsourcing, pilot tests were found important to make sure that we were gathering the right information. Studies prior to the main task is an important step. Applying a good user interface design in the crowdsourcing tasks might help to make sure that the tasks are easier to understand.

A few questions that can be asked when designing the tasks:

- *Are the instructions clear?*

  This is important to make sure the crowdsourcing participants fully understand and know what to expect from the tasks. Note that the crowdsourcing participants have different backgrounds, so, a clear explanation on what they have to do is vital. Thus, the pilot test is one way to help us make sure the participants know what to do and are able to give appropriate answers.

- *Is the task simple and easy to follow?*

  Most participants in crowdsourcing platform are looking for jobs that are easy and quick to complete. Avoiding complexity in a job is preferred. The quality of the tasks should also be considered. In CrowdFlower, we can set the performance level (Figure 5.11), where the tasks should be balanced between speed and quality.

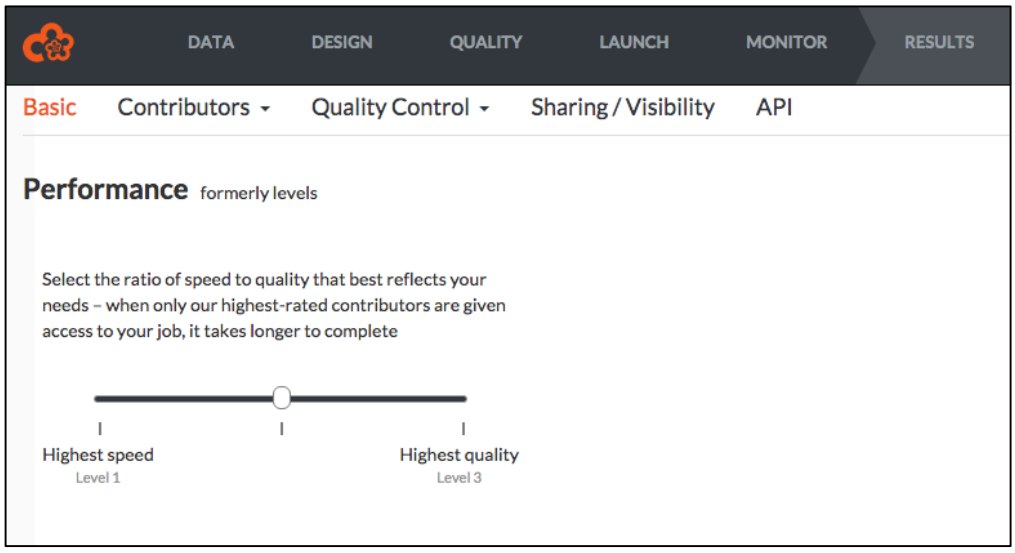


Figure 5.11: Screenshot of the Performance Level setting

- *Are we gathering the information that we want?*

  In the pilot test, the most important part is to analyse the results to ensure that the results we obtain are what we expect. Based on the results, we can also ensure that our settings would

give us an achievable task. For example, we can exclude or include participants from certain regions/countries that we think would give us the best results for our tasks (Figure 5.12).

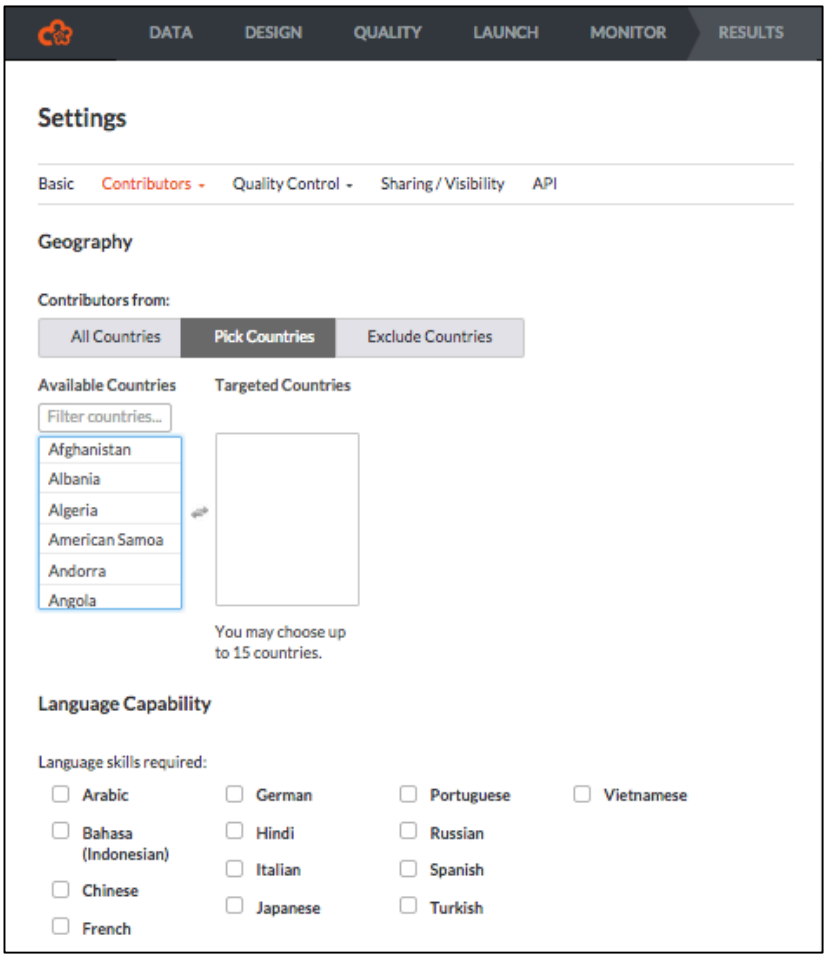


Figure 5.12: Participants (Contributors) settings

Another important feature in CrowdFlower is the test question, as discussed in Section 5.2.2. This serves as a quality control on participants, to make sure that they are doing the tasks as given. If a participant fails to answer the test question too many times (level of accuracy is low), CrowdFlower would automatically stop the job. However, this would also affect the minimum total payment (agreed upon assigning the job) and trigger an automatic email informing the job owner that the job has run out of funds. It is also important to set the maximum judgment per contributor as this feature is used

to restrict the participants doing the task too many times that they can view the test questions more than once.

We discovered that it is essential to have a clear idea on what we want from the crowdsourcing platform. It is advisable to go through the settings, not just using the default. CrowdFlower have powerful tools in the settings to make sure the jobs are done in the right way.

5.8 Discussion

Since our ROUGE scores did not provide significant differences for different tweet-biased settings, we believed that there would be other ways to evaluate our summaries. Mackie et al. (2014) and Yulianti et al. (2015) explored the use of crowdsourcing to evaluate microblog summaries. Based on the results showed in Table 5.1, all of the summaries generated using related tweets (EXP, EXP_s, T10.EXP, and LD.EXP) outperformed the local document summaries.

We found that the number of tweets did not impact summary generation. The paired summaries of EXP-T10.EXP (T10 used only the ten best tweets) showed that almost 60% of the participants agreed that the summaries generated were the same. Here we believed that a small number of related tweets could help to identify important information of the document and thus, extract the same sentences (when compared with summaries generated using more tweets). We also believed that a large number of tweets contained the same information, where most of the tweets are re-posting of the original tweet (Retweet). Thus, this does not give a big impact on the sentence selection. We assumed that this might be the reason for a small gap of the Recall and Precision score between EXP-T10.EXP; with score differences of 0.005 (Recall) and 0.007 (Precision), which showed no significant difference between the two settings.

To compare the human judgement and the ROUGE score results, we observed the ordering of both evaluations. Based on the Precision scores in both ROUGE-1 and ROUGE-2, EXP showed a

better score compared to all settings and significantly better than EXP_s. Thus, we ranked our EXP to be better than EXP_s based on the precision score.

For our participant preference, we ordered summary settings based on the Condorcet ranking method as showed in Table 5.4, where EXP wins in the paired voting system. We observed that LD.EXP is preferred to EXP_s, and LOCAL is the least preferred. The findings from both evaluations showed that the human judgements have the same first rank order with our ROUGE score (EXP).
We also paired the automated generated summaries with the BASELINE used in Chapter 4. Here we can see that BASELINE was preferred by participants, except when paired with EXP. Using the Condorcet and Ranked Pairs method, EXP was ranked number one when we compared with all the other settings.

In general, it is difficult to compare and report the correlation between automated and human judgement evaluation since there is little work on human judgement evaluation reported in document summarization. Mackie et al. (2014) describe a similar experiment where they summarize 135 tweets in 50 topics using three different systems over different datasets, including the TREC 2011 Microblog track dataset. However, they did not report ROUGE result for the Microblog dataset as there were no gold standard summaries provided.

For their Recall and Precision ROUGE-1 evaluation, they applied it only to one dataset (the *trending-topics-2010* from Sharifi et al. (2010)). Their results showed that there was a significant difference for one system only (system ranked at number 1). They also reported that the Recall ROUGE-1 metric showed the same ranking of human judgement for the generated summaries. For the Precision score (which showed a significant difference in all three systems), they highlighted that their participant ranking was different from their Precision ranking. They believed that this is because the reference summary for their automated evaluation might not contain all the key information of a document.

In the qualitative analysis, we were interested to examine the reasons for participant choices by gathering and categorizing comments into six categories: Presentation, Topic Discussed,

Document-Summary Similarity, Same Summary, Not Classifiable and Spam. From our analysis, we found that there are comments that compare the similarity between the local document and the chosen summary (Document-Summary Similarity). As discussed in de Oliveira (2005), we assumed that the quality of a summary can be measured if it has similarity with the original document. We found similar categories of Content and Presentation in work by Mackie et al. (2014) and Yulianti et al. (2015), however, we believed that our categories are more specific to the comments based on our paired summaries.

We can see from Figure 5.6 - 5.9 that the content of a summary is important. The topic of the summary (Topic Discussed) and the similarity between the original document and its summary (Document-Summary Similarity) are the most given reason for preferring a summary. Our results showed that 48.4% of the given reason was in the Topic Discussed category and 38.7% was in Document-Summary Similarity. This was also shown in work done by Mackie et al. (2014) and Yulianti et al. (2015) where they also identified that the content and the topic discussed in the summary was the main reason for participants choice. Mackie et al. (2014) showed 55.8% (merging the number of informativeness, readability and tweet-related) as the topic-related (content) for choosing a summary, whereas Yulianti et al. (2015) showed 92.3% of participants stated that Content was the main reason to prefer a summary. Note, Yulianti et al. (2015) participant comments into three categories with two focussed on presentation.

The Presentation category showed only a small percentage in all settings (12.8%). Thus, when comparing two summaries, only a small number of participants believed that the structure and order of the sentences were important when choosing one, which is also showed by Yulianti et al. (2015. However, Mackie et al. (2014) has a category: Readability and Length, which participants choose 44.2% of the time. We believed that Readability is an important aspect: summaries need to be understandable.

The findings that the summaries generated by EXP and T10.EXP have similar extracted sentences may be the reason for insignificant ROUGE scores between the two Affinity Graph settings.

Examining these comments also showed that the number of tweets did not have any significant effect on the summaries. We also see that EXP and EXP_S showed that it can extract better sentences that are topic-related; and the summaries generated by T10.EXP and LDEXP showed similarities with the local document.

We also compared our summaries with another tweet-biased summarization system (TBS_{sn}) Yulianti (2013) and Yulianti et al. (2015). TBS_{sn} was no different to EXP for the Recall ROUGE-1 evaluation but was favoured by participants. In the latter result, participant comments 54.5% of the time were ambiguous. Participants choose TBS_{sn} just because "*it is better*". However, the results in Table 5.8 showed that TBS_{sn} was less elected as the best summary when paired with BASELINE summaries. We believed that the summaries generated by TBS_{sn} might be similar to BASELINE summaries, as showed in Figure 5.10. This also might explain why TBS_{sn}'s ROUGE score was higher and has small difference compared with BASELINE.

5.9 Conclusion

The findings and discussions of the CrowdFlower experiment allowed us to answer the following research questions. In the human judgement results, all of the tweet-biased Affinity Graph settings were more preferable. The Condorcet voting system showed that the tweet-biased Affinity Graph summaries are ranked higher (1^{st} to 3^{rd} place) than the LOCAL summaries. EXP was ranked better than BASELINE and TBS_{sn} summaries, even though both of the latter approaches gave better ROUGE scores compared to EXP. Both results answered our RQ4, where the tweet-biased Affinity Graph approaches could improve single document summaries.

The findings for RQ4 helped us to answer the third research question (RQ3), where we assumed that human judgement approach could give us a better and reliable result compared to the ROUGE score discussed in Chapter 4. Based on the human judgement results, we found that the

human judgement gave significant results in evaluating the summaries. We also report the findings from the comments, which provided new insights in identifying features to auto-generate a summary.

Thus the contributions from the chapter were:

1. A methodology to develop a CrowdFlower pair-wise summary evaluation.
2. We found our tweet-biased Affinity Graph approaches were favoured by participants. Related tweets provided valuable information to summarisers.
3. We identified features participants mainly used to choose their preferred summaries: topic discussed and document-summary similarity.

We also describe lessons learnt from the CrowdFlower experiments. We believed that human judgement evaluation using crowdsourcing platform could provide a reliable and fast result for document summarization evaluation. More work is needed to give better insight into the use of crowdsourcing platform for summarization and also to learn how results correlate with the automatic evaluation.

Chapter 6
Conclusion

We set out to investigate an approach to summarize a single document in a multi-document environment and as well as an evaluation method. Our work discussed the development of a summarization system, a new dataset to test the effectiveness of a summarization approach and also proposed an alternate methodology to evaluate the summaries.

From our literature review, we see that most summarization research implemented either a single or a multi-document approach. We hypothesized that a single document summary could be improved using information from related documents. The related documents carry important information that can be exploited to improve summarization accuracy. To this end, we discussed the use of an Affinity Graph that draws in related documents including social media content, such as tweets.

We also explored summary evaluation. We believed that the standard automated evaluation used by many researchers in document summarization are hard to implement. Different settings (especially in ROUGE) need to be tested in order to find the best parameter for a specific summarization system. Even the methodology to produce a gold standard (or reference) summary is time-consuming and costly. Thus, a new evaluation that is more user-driven should be explored.

6.1 Thesis Contribution

The contributions of the thesis are presented by discussing each research question (RQ) in turn.

RQ1: How Effective are Graph-Based Algorithm Approaches to Improve Single Document Summarization?

First, the thesis discussed the use of an Affinity Graph to summarize single document by including related information from surrounding documents. Here, we focused on the exploration of the parameter space (different similarity measures, the number of related documents and length of a document) of the Affinity Graph algorithm. The main reason to do this is to test the robustness of the approach and investigate if there are certain settings to be applied to get the best result from the approach.

We tested the algorithm using the same dataset reported in the previous work. Similarly, we found that the Affinity Graph summaries that used information from other documents improved significantly the summaries that derived from local documents only. However, our findings showed that the use of different parameters gave equally good ROUGE results.

We also noticed that the summaries generated using Lead Paragraph documents produced summaries that are similar with the full document summaries. We believed that this setting, where a limited amount of related documents are used, could be further investigated. Hence, this leads to the second set of experiments for the thesis.

RQ2: Can the Affinity Graph Algorithm Improve Single Document Summarization when Using Limited Length Documents (e.g.: tweets)?

The second discussion in the thesis has three parts: the development of an evaluation dataset; a new Affinity Graph framework that includes related tweets; and the discussion on the results, including the manual comparison of the summaries produced by different Affinity Graph settings.

In the first part, we discussed the development of a new dataset based on a tweet collection from TREC 2011's Microblog track dataset. We identified web documents related to a set of tweets. We focussed on documents had at least 10 tweets pointing to them. We applied a ROUGE evaluation; however, since there were no gold standard summaries to be used as reference summaries, we had to develop a system to manually extract important and relevant sentences from web documents. Here, we got help from postgraduate students and university staff.

In the second part, we defined the tweets as the 'limited' information in Affinity Graph approach and tested our summariser on the evaluation data set.

In the third part, we discussed the findings from the new Affinity Graph approach. The summaries were generated from four different settings of the tweet-biased Affinity Graph approach (namely EXP, EXP_s, T10.EXP, and LD.EXP). However, they did not produce a better ROUGE score as we expected; this was due to the small dataset used in the experiments. We also believed that the methodology to generate the gold standard summaries could be improved, so it would generate better summaries as references for the ROUGE evaluation.

When analyzing the findings, we noticed that for both datasets, there is a negative correlation between document length and the ROUGE scores: longer documents tend to produce lower Recall ROUGE scores. Thus, we believed more tests should be done to different datasets to conclude if such relationship is true. In a tweet-ROUGE analysis, we discovered that the number of tweets pointing to a document had no effect on ROUGE scores.

In the analysis, we also noticed that inclusion of the expanded tweets improved summary accuracy. This can be seen in the randomly selected documents with their summaries in Figure 4.12 – 4.15. The more related tweets there were, the more relevant sentences were extracted for the summaries, as in Figure 4.12 and Figure 4.14. We believed that a maximum number of tweets would be able to add more information in generating a good summary, especially for long documents that have many topics.

Generally, the tweet-biased summaries extracted sentences that are more topic-focused. This is because most related tweets contain the title of the document or words/phrases that is considered 'important' by the tweets' users. We would suggest the tweet-biased Affinity Graph approach be used for documents with longer length (200 sentences or more) and have a high number of tweets pointing to them.

RQ3: Is a Crowdsourced Human Judgement Approach a Better Evaluation Compared to the Standard Automated Summary Evaluation?

From the findings in research questions one and two, we believed that the Affinity Graph approach is able to create good summaries, but we failed to get significant results in our ROUGE evaluation. This is due to few reasons:

1. We created only a small sample dataset: 10% of the whole collection. While this scale of collection was predicted to be sufficiently large, our post-hoc power analysis showed that the number of samples used was too small to reject a Type II error.
2. A problem that we encountered when implementing the automated evaluation was to find 'experts' that would have the time to read, understand, and summarize the documents. In DUC02, the experts (human) summarizers were asked to write an abstract summary for all of the documents. But in our SESys, we asked our volunteers to select 'the most relevant' sentences (to do *extractive* summaries, not *abstractive*). This may cause difficulties for the volunteers in identifying sentences to summarize the document, as (long) documents may have more than one topic.

Because of these reasons, we assumed that our reference summaries might not be ideal for evaluation. In our summary analysis, we found that a great deal of information was captured in the tweet-biased summaries. Thus, it is in our attention to find an alternate way to evaluate our automated

summaries. Based on the current literature in document summarization, we found that a user preference approach (which has successfully applied in relevance judgments and text annotation studies) is a promising way to analyze the summaries.

Via CrowdFlower, we asked participants to choose their preference in a series of pairwise comparisons of different tweet-biased Affinity Graph settings. We found that the user preference approach could give us better and more reliable results compared to our ROUGE-based collection. Additionally, comments from CrowdFlower workers gave us additional information on summary quality. We were able to identify features based on the comments and group them into five categories. The findings are consistent with findings from past studies, which found that the *Content* of the summary is the most important feature when choosing a preferred summary. However, in contrast with previous studies, we split the content variables into two smaller groups, Topic Discussed and Document-Summary Similarity. We believed that this finding could be further used to develop better-automated summarization techniques.

We also discussed the lessons learnt from the CrowdFlower experiments. We hoped that the discussion would be able to give new insight and create an opportunity to improve the methodology of user preference experiments. We hope more work in document summarization would consider using this evaluation method.

RQ4: Will the tweet-biased Affinity Graph approach be preferred over LOCAL settings?

From our user-preference study, it was found that the tweet-biased Affinity Graph settings were more preferable compared to the LOCAL settings. This is based on the Condorcet voting system that showed the summaries generated using the Expanded document are ranked first in a Condorcet matrix. The same result is also revealed in the manual comparison of the summaries, where the tweet-biased

Affinity Graph summaries were preferred, especially for long documents. This is in line with our hypothesis that related tweets provide relevant information when summarizing a local document.

6.2 Future Work

Our findings suggest that it is possible to incorporate related documents to improve single document summaries, where we explore the use of Affinity Graph in a multi-document environment. Future work could investigate the range of parameter settings in an attempt to improve document summarization accuracy.

In a further analysis of the documents and its related tweets, we noticed that most are retweets and/or only contain Twitter username mentions (@username). This is because most of the tweets are reposts of an original to share information, a URL, or a web document. More processing could be done to the tweets, such as clustering before we use the tweets to improve summaries. Such processing should summary accuracy.

It is recommended that further research be undertaken in the document length-ROUGE score relationship, as ROUGE scores should not be penalized for longer documents. A number of possible future user preference studies could be undertaken. It would be interesting to see if there is any relationship between document length and user preference, as this would be comparable with the document length-ROUGE score relationship.

One could also explore the relationship/correlation between user judgement and ROUGE score. Past work by Dorr, et.al (2004) found that there is a small but significant correlation between intrinsic and extrinsic evaluation. We believe a future study to investigate this would be beneficial, as this would establish if user preference evaluation is equally reliable with a ROUGE-based summary evaluation.

Appendix A

A.1 LEMUR PROJECT TOOLKIT

The Lemur Project was developed by the Center for Intelligent Information Retrieval (CIIR), University of Massachussets, Amherst and the Language Technologies Institute (LTI), Carnegie Mellon University. The Lemur Project supports the use of statistical language models, especially for information retrieval tasks. The toolkit also includes the Indri Search Engine, where it is used for large-scale search.

For the project, Indri 5.5 and Lemur 4.12 are used for similarity search for our document and sentence similarity (Part 1a and 1b in Figure 3.2, Chapter 3).

INDRI SEARCH ENGINE

INDRI Search engine was developed as one component in the LEMUR Toolkit (Strohman et al., 2005). It has two main components: the query language and retrieval model, where both support retrieval at different level and type.

In the INDRI search engine, there are two main functions:

1. **IndriBuildIndex**

   This function is to build an index for the dataset that we used in our experiment. For each of the document in the dataset, INDRI will build an index for the query retrieval. To run

the **IndriBuildIndex**, we need to prepare a parameter file (in XML format) for all documents as follows:

```
<parameters>
<index> path_to_index_repository </index>
<corpus>
  <path>path_to_corpus </path>
  <class> define_file_format </class>
</corpus>
<memory> define_memory_value </memory>
<stemmer>
  <name>define_stemmer </name>
</stemmer>
<metadata> define_metadata_field_(header/title) </metadata>
<stopper>
  <path>path_to_stopwordlist </path>
</stopper>
</parameters>
```

*Figure 1: Parameter file format for **IndriBuildIndex***

The parameter file will create an index file for the dataset, and it will be used to query the document. Here we created two different index files for Document Similarity (to index the document) and Sentence Similarity (to index each sentence from the document).

2. IndriRunQuery

This function is used to query the index file that we have created. For each of the query, we need to create a query file as the format below:

```
<parameter>
<query>
  <number> query_number </number>
    <text> query_text </text>
</query>
```

Figure 2: Query file format

Here we used the Local Document as the query for the **IndriRunQuery** function. For this, we created two different files for document-document similarity (Figure 3) and sentence-sentence similarity (Figure 4).

```
<parameter>
<query>
  <number>AP880911-0016</number>
  <text>Hurricane Gilbert swept toward the Dominican Republic Sunday and the
Civil Defense alerted its heavily populated south coast to prepare for high
winds heavy rains and high seas The storm was approaching from the southeast
with sustained winds of 75 mph gusting to 92 mph There is no need for alarm
Civil Defense Director Eugenio Cabral said in a television alert shortly before
midnight Saturday
</text>
</query>
```

Figure 3: Query for Document Similarity

```
<parameter>
<query>
  <number>AP880911-0016_1</number>
  <text>Hurricane Gilbert swept toward the Dominican Republic Sunday and the
Civil Defense alerted its heavily populated south coast to prepare for high
winds heavy rains and high seas</text>
</query>

<query>
  <number>AP880911-0016_2</number>
  <text>The storm was approaching from the southeast with sustained winds of 75
mph gusting to 92 mph</text>
</query>

<query>
  <number>AP880911-0016_3</number>
  <text>There is no need for alarm Civil Defense Director Eugenio Cabral said
in a television alert shortly before midnight Saturday</text>
</query>

<query>
  <number>AP880911-0016_4</number>
  <text>Cabral said residents of the province of Barahona should closely follow
Gilberts movement</text>
</query>
</parameter>
```

Figure 4: Query for Sentence Similarity

For each of the document in the dataset, we had done pre-processing where we remove all symbols (.,/$!") and split the document into sentences. Once we have prepared the query file and indexed for the dataset, we run the **IndriRunQuery** as follows:

:indri-5.0/runquery/IndriRunQuery=query_parameter_file -count=number_of_result* -index=/path/to/index -trecFormat=true > *result_file***

The **IndriRunQuery** function will generate an output file (result – Figure 4), where the similarity values between the document-document and sentence-sentence relationship will be produced. Figure 5 showed an example from the DUC documents and the columns represents:

<queryID> Q0 <DocID> <rank> <score> <runID>

where:

(1) **<queryID>** is the Local Document or sentences from the Local Document to be summarized (e.g.: AP880911-0016_1 is sentence 1 from document AP880911-0016).

(2) **<DocID>** is the related documents or sentences from the related documents to be summarized (e.g.: AP880915-0003_32 is sentence 32 from document AP880915-0003).

(3) **<rank>** is the rank for each sentence from the related documents.

(4) **<score>** is the similarity scores of the sentences from the local document to the sentences from the related documents.

(5) **<runID>** is the run name for the query (Exp represents the Expanded settings)

```
AP880911-0016_1 Q0 AP880915-0003_32 1 0.394769 Exp
AP880911-0016_1 Q0 AP880912-0137_10 2 0.277971 Exp
AP880911-0016_1 Q0 AP880916-0060_12 3 0.239221 Exp
AP880911-0016_1 Q0 AP880916-0060_11 4 0.220191 Exp
AP880911-0016_1 Q0 AP880916-0025_55 5 0.217282 Exp
AP880911-0016_1 Q0 AP880915-0142_38 6 0.217282 Exp
AP880911-0016_1 Q0 AP880915-0003_34 7 0.217282 Exp
AP880911-0016_1 Q0 AP880914-0131_44 8 0.217282 Exp
AP880911-0016_1 Q0 AP880915-0003_31 9 0.203774 Exp
AP880911-0016_1 Q0 AP880916-0025_7 10 0.203332 Exp
AP880911-0016_2 Q0 AP880912-0095_32 1 0.371327 Exp
AP880911-0016_2 Q0 AP880912-0095_29 2 0.209033 Exp
AP880911-0016_2 Q0 AP880914-0131_39 3 0.169542 Exp
AP880911-0016_2 Q0 AP880915-0003_45 4 0.157445 Exp
AP880911-0016_2 Q0 AP880915-0142_31 5 0.15652 Exp
AP880911-0016_2 Q0 AP880915-0142_32 6 0.13354 Exp
AP880911-0016_2 Q0 AP880916-0060_9 7 0.125609 Exp
AP880911-0016_2 Q0 WSJ880912-0064_14 8 0.106718 Exp
AP880911-0016_2 Q0 AP880916-0025_50 9 0.104148 Exp
AP880911-0016_2 Q0 AP880915-0142_34 10 0.104148 Exp
```

Figure 5: INDRI Output File (Result) with the similarity scores

These similarity scores will be the input to our AG algorithm.

LEMUR TOOLKIT

We used LEMUR Toolkit to calculate the document-document and sentence-sentence similarity using Cosine Similarity. There are four main steps for LEMUR toolkit:

1) Prepare the QUERY document

Similar with INDRI, we must prepare the query file for each document and sentence set before we can run LEMUR to calculate the Cosine Similarity. The format for LEMUR query file is the same with INDRI, as in Figure 3 and Figure 4.

2) **To build the document index: BuildIndex**

This function is a similar function as **IndriBuildIndex** in INDRI (to build an index). However, we need to prepare a list of files for the document dataset (to be called in the <dataFiles>) as in Figure 6:

```
<parameters>
  <index> key_file_in_Index </index>
  <indexType> key </indexType>
  <memory> define_memory_value </memory>
  <docFormat> trec </docFormat>
  <position> true </position>
  <stemmer> define_stemmer </stemmer>
  <stopper> define_stopword_list </stopper>
  <dataFiles>.lst_files </dataFiles>
</parameters>
```

*Figure 6: Parameter file format for **BuildIndex***

This function will create an index for the assigned document/dataset. The index is stored in the <index> path as defined in the parameter and will be called in the **RetEval** function. To run the **BuildIndex** function, we need to call it in LEMUR:

:lemur-4.12 $ BuildIndex *parameter_file*

3) Parse the query document: ParseToFile

For the query in LEMUR, we need to prepare a parameter file to parse the document/sentence. This function will create an out file, which will store all the words from the document and sentence query. We need to parse two different query file for document-document dataset and sentence-sentence dataset.

```
<parameters>
  <docFormat> format </docFormat>
  <outputFile> path_to_out.file </outputFile>
  <stemmer> define_stemmer </stemmer>
  <stopwords> define_stopword_list </stopwords>
</parameters>
```

*Figure 7: Parameter file format for **ParseToFile***

To run the function, we need to call it in LEMUR:

:lemur-4.12 $ ParseToFile ***query_parameter_file query_file***

The *query_ parameter_ file* is defined as in Figure 7, and the *query_ file* is as defined in Figure 3 and 4. Examples of the out file are shown in Figure 8.

```
<DOC AP880911-0016>
hurrican
gilbert
head
dominican
coast
hurrican
gilbert
swept
dominican
republ
sundai
civil
defens
alert
heavili
popul
south
coast
prepar
high
wind
heavi
rain
high
sea
storm
approach
southeast
sustain
wind
75
mph
gust
92
mph
</DOC>
```

```
<DOC AP880911-0016_1>
hurrican
gilbert
head
dominican
coast
hurrican
gilbert
swept
dominican
republ
sundai
civil
defens
alert
heavili
popul
south
coast
prepar
high
wind
heavi
rain
high
sea
</DOC>
<DOC AP880911-0016_2>
storm
approach
southeast
sustain
wind
75
mph
gust
92
mph
</DOC>
```

Figure 8: out file for Document dataset (left) and Sentence dataset (right)

4) Run the retrieval model: RetEval

The final step in LEMUR is to run the **RetEval** function. This function will generate the results for the Cosine Similarity retrieval model. For this function, we will need to create another parameter file as in Figure 9.

```
<parameters>
  <index> index_path </index>
  <retModel> define_retrieval_model </retModel>      // 0 for TF-IDF,
                                                     // 1 for Okapi,
                                                     // 2 for KL-divergence,
                                                     // 5 for cosine similarity
  <textQuery out.file_path </textQuery>
  <resultCount> number_of_result </resultCount>
  <resultFile> result.file_path </resultFile>
</parameters>
```

*Figure 9: Parameter file format for **RetEval***

To run the function, we need to call it in LEMUR:

:lemur-4.12 $ RetEval ***query_Retrival_file***

The **RetEval** function will generate an output file (result) with the same format as the **IndriRunQuery**. These similarity values will be the input to our Affinity Graph algorithm.

A2. AFFINITY GRAPH SETUP

The results as in Figure 5 will be used as the input for the Affinity Graph algorithm. An Affinity Graph summarization system was developed using PHP.

(1) Read the output file from LEMUR/INDRI
(2) Create the matrix M using Equation 3.1
    IF sentence from the query_doc THEN $\lambda = 1$
    ELSE $\lambda \times sim_{sen}(s_i, s_j)$
(3) Normalize the matrix M using Equation 3.2 (output in Figure 11)
(4) Calculate the *if_score* using Equation 3.3
(5) Sort the sentences based on the *if-score* (output in Figure 12)
(6) Identify the sentences based on the sentence_id
(7) Count the summary word and truncate the summary when reach 100 words (output in Figure 13)

Figure 10: Affinity Graph algorithm

From the output file from INDRI/LEMUR, only the <queryID> <DocID> and <score> will be used for the next step in Affinity Graph. For each of the score, a matrix will be created based on Equation 3.1 and then normalized using Equation 3.2. Thus, the normalized matrix will create an output as in Figure 11.

```
AP880911-0016_1 AP880915-0003_32 0.085842913819 0.00308367414876
AP880911-0016_1 AP880912-0137_10 0.069667037817 0.00250259962039
AP880911-0016_1 AP880916-0060_12 0.052994867351 0.00190369705776
AP880911-0016_1 AP880916-0060_11 0.048779132421 0.00175225819993
AP880911-0016_1 AP880916-0025_55 0.063856355134 0.00229386658491
AP880911-0016_1 AP880915-0142_38 0.062427508702 0.00224253914728
AP880911-0016_1 AP880915-0003_34 0.047248188182 0.00169726317515
AP880911-0016_1 AP880914-0131_44 0.062689550794 0.00225195229962
AP880911-0016_1 AP880915-0003_31 0.044310860074 0.00159174761947
AP880911-0016_1 AP880916-0025_7 0.059756631484 0.00214659511806
AP880911-0016_2 AP880912-0095_32 0.134665078493 0.00483747816581
AP880911-0016_2 AP880912-0095_29 0.075807698747 0.00272318623056
AP880911-0016_2 AP880914-0131_39 0.048915749214 0.00175716578816
AP880911-0016_2 AP880915-0003_45 0.034236572695 0.00122985613448
AP880911-0016_2 AP880915-0142_31 0.04496991772 0.00161542248015
AP880911-0016_2 AP880915-0142_32 0.03836751094 0.0013782489011
AP880911-0016_2 AP880916-0060_9 0.027826287379 0.000999583998598
AP880911-0016_2 WSJ880912-0064_14 0.034948971102 0.00125544711751
AP880911-0016_2 AP880916-0025_50 0.030607743276 0.00109950026733
AP880911-0016_2 AP880915-0142_34 0.029922866028 0.00107489790738
AP880911-0016_2 AP880915-0142_3 0.02907300009 0.00104436877566
```

Figure 11: The output for normalized matrix where the column represents (1) The local document sentences, (2) The related document sentences, (3) The matrix value from Eq. 3.1 and (4) The normalized value from Eq. 3.2

Based from the normalized matrix values, the *if_score* is computed using Equation 3.3. Once all the *if-score* for all the sentences have been calculated (we add it up for each local document sentences), it is then sorted and the output is shown in Figure 12.

```
0.000451307726875 AP880911-0016_07
0.000448230213961 AP880911-0016_06
0.000440045050582 AP880911-0016_13
0.000436998839210 AP880911-0016_08
0.000435402371009 AP880911-0016_01
0.000434945847498 AP880911-0016_09
0.000433444094522 AP880911-0016_05
0.000429267042264 AP880911-0016_12
0.000428096355953 AP880911-0016_11
0.000425322662954 AP880911-0016_14
0.000425098815735 AP880911-0016_15
0.000424855191204 AP880911-0016_04
0.000417724537160 AP880911-0016_10
0.000416925762445 AP880911-0016_03
0.000414595839098 AP880911-0016_02
```

Figure 12: The sorted if-score for all sentences

In the results (Figure 12), we can see that sentence number 7 for document AP880911-0016 has the highest score, thus will be the first sentence for the summary. Based on the sentence and document number, the sentences will be selected from the local documents. The last step is to count the words for the summaries and truncate the summaries to 100 words (Figure 13).

```
[1] The National Hurricane Center in Miami reported its position at 2
a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140
miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo
Domingo.
[2] Tropical Storm Gilbert formed in the eastern Caribbean and
strengthened into a hurricane Saturday night.
[3] On Saturday, Hurricane Florence was downgraded to a tropical storm
and its remnants pushed inland from the U.S. Gulf Coast.
[4] The National Weather Service in San Juan, Puerto Rico, said
Gilbert was moving westward at 15 mph with a ``broad area of
cloudiness and heavy weather'' rotating around the center …
```

Figure 13: Document AP880911-0016 Hurricane Gilbert Heads Toward Dominican Coast

Appendix B

B.1 MEDIUM-LENGTH DOCUMENT

Hawks & Handsaws: A few blunt words

31 JUL 93 | By MICHAEL THOMPSON-NOEL

THE revelation that John Major is capable of candid, blunt and salty language when talking off-the-record to friendly journalists has surprised some people. It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good.

With luck, it is reckoned, Major's image as a leaden-tongued wimp may undergo correction.

What piffle. Major is a gonner, especially after this week's revolt of the wooden-tops in the Christchurch by-election, where a Conservative majority of 23,015 at last year's general election was converted into a 16,427 majority for the Liberal Democrats. Fifteen months too late, the voters of Christchurch rounded on the Tories with a malignant and squeaky fury.

In reality, all politicians, not just Major, are far more candid and salty when chatting in private than when speaking in public. In public, they have to be careful of what they say, so their utterances achieve a horrible mattness. But in private they relax. Their syntax disappears. Their words become nonsensical. They swear and joke and shout. It really is a spectacle.

To show you what I mean, I spoke yesterday to John Major and John Smith. Smith, a Scot, is leader of the Labour Party, though not many people know that. In the aftermath of Christchurch, where Labour lost its deposit, I wanted to provoke the two Johns into a spot of real soul-searching.

To guarantee them privacy, I used a signal-scrambler. No one could have eavesdropped. Their responses were true to form. But I have left out the swear-words because the new Financial Times Style Guide states that 'the gratuitous use of expletives or obscenities is discouraged . . . Four-letter

expletives will usually be confined to infrequent use in the review (Arts) pages.' I can live with that, though why the artsy-fartsies should receive any dispensation is a puzzle.

First, I tackled Major. I said: 'Did you read, John, what Olivier Blanchard, Rudiger Dornbusch, Stanley Fischer, Franco Modigliani, Paul A Samuelson and Robert Solow wrote, in just one article, in the FT this week? They were describing Europe's lunatic monetary policies and exchange rate arrangements. They did not pull their punches. I bet you went chalk-white.

'So why not walk the plank, John? You are the most unpopular prime minister since the start of the fourth century. Why invite more punishment? Unfairly or not, you are drawing the blame for all life's unpleasantnesses, let alone the cock-ups.'

'Are you sure?' the prime minister replied. 'I mean . . . how did it come about, Michael . . . like, Christchurch, y'know - load of . . . let me put it to you - the economy, of course . . . I mean, wimpy guy like me. But I'm not giving in like that, like . . .'

On and on it went. Then I rang John Smith. I told him I had been impressed with his interview with Andrew Marr in The Independent on Thursday, in which he sharpened up his promise to introduce meaty political reforms (if he ever gets elected), including a referendum on proportional representation.

I said: 'You are starting to raise your game, John. Many people will have agreed with your assertion that democracy in Britain is decaying, and that the Tories must be roasted for their arrogance, incompetence, complacency and sharp practices - especially their 'centralisation of power and the

elimination of opposition'. But some of your critics still accuse you, John, of laziness and ineffectualness. What do you say to that?'

'Away, ye thowless jad,' shouted the Labour leader. 'Gie me o'wit an' sense a life, behint a kist to lie an' sklent. Our Stibble-rig was Rab M'Graen, a clever, sturdy fellow, but then he was sae fley'd by his showther gae a keek, an' tumbl'd wi' a wintle. Likewise with political and constitutional reform, Michael, for by the L - - d, tho' I should beg wi'lyart pow, I'll laugh, an' sing, an' shake my leg, as lang's I dow]'

After that, I thought of telephoning Wing-Commander Paddy Ashdown, leader of the Liberal Democrats, to solicit his views on Christchurch. But I couldn't raise the energy.

Appendix C

C.1 LONG-LENGTH DOCUMENT

Leonard Bernstein Dies; Conductor: Composer: Music: Renaissance man of his art was 72. The longtime leader of the N.Y. Philharmonic carved a niche in history with 'West Side Story.'

October 15, 1990 | MYRNA OLIVER | TIMES STAFF WRITER

Leonard Bernstein, the Renaissance man of music who excelled as pianist, composer, conductor and teacher and was, as well, the flamboyant ringmaster of his own nonstop circus, died Sunday in his Manhattan apartment. He was 72.

Bernstein, known and beloved by the world as "Lenny," died at 6:15 p.m. in the presence of his son, Alexander, and physician, Kevin M. Cahill, who said the cause of death was complications of progressive lung failure. On Cahill's advice, the conductor had announced Tuesday that he would retire. Cahill said progressive emphysema complicated by a pleural tumor and a series of lung infections had left Bernstein too weak to continue working.

In recent months, Bernstein canceled performances with increasing frequency. His last conducting appearance was at Tanglewood, Mass., on Aug. 19.

Bernstein was the first American-born conductor to lead a major symphony orchestra, often joining his New York Philharmonic in playing his own pieces, while conducting from the piano.

He etched other niches in history by composing the indelible "West Side Story" and teaching a generation about classical music via the innovative television series "Omnibus."

Exhibiting remarkable talent and expertise in four areas that most artists wish they possessed in merely one, Bernstein still might have remained an obscure musician without the unique theatrical flair that dominated his personal as well as professional life. With it, he became a *personality* , well

known even to people who never bought a ticket to a musical performance or watched a serious television show.

The dervish persona, including his upstart gymnastics on the podium, never lessened throughout his long life in the spotlight.

He made classical music understandable and palatable to the masses. And he lifted popular music to a higher plane, infusing performers and listeners with his manic joy in creating tonal sound.

"Some conductors mellow with age," commented Times music critic Martin Bernheimer when Bernstein conducted the Los Angeles Philharmonic at UCLA in 1986. "(But) Bernstein, at 68, remains a frenetic combination of orbiting rocket, aerobics master, super-juggler, matinee idol, booming cannon, hysterical mime, heart-rending tragedian, bouncing ball, sky writer, riveting machine, mawkish sentimentalist and danseur ignoble."

Describing the conductor in the same concert, Bernheimer referred to him as "the shrugging, jumping, sighing, soaring, gushing, crouching, rocking, rolling, bounding, bobbing, leaping, jiggling, stabbing, hunching, bumping, grinding and grunting maestro in excelsis."

Critics also were quick to agree that had his envied and often-criticized showmanship masked lazy, sloppy or inept musicianship, Bernstein could never have remained an internationally sought-after conductor for five decades. He knew what he was doing, and the musicians he accompanied, wrote for, conducted, or lectured to and taught admired him as one of their own.

Louis Bernstein (so-named because his maternal grandmother insisted) was born Aug. 25, 1918, in Lawrence, Mass., to two Russian Jewish immigrants. His father, Samuel Joseph Bernstein, was an entrepreneur of women's hair care products and a Talmudic scholar. His mother, Jennie Resnick Bernstein, who survives him, said her son always had an ear for music. "When he was 4 or 5, he would play an imaginary piano on his windowsill."

The parents preferred the name "Leonard" and called the boy that. When his kindergarten teacher asked "Louis Bernstein" to stand up, he remained seated and looked around the room to see who shared his last name. Bernstein changed his name legally at age 16, when he got his first driver's license.

His mega musical talent emerged belatedly and almost by accident.

When Bernstein was 10, a divorcing aunt stored her old upright piano with his parents, and the boy who used to play at the windowsill became fascinated with it. He asked for lessons, and soon was playing better than his teacher, a neighbor's daughter who charged $1 a lesson.

By age 12, he was studying at the New England Conservatory of Music and had determined, despite his father's objections, that music--at that point playing the piano--would be his career.

Bernstein's stunning instinctive talents for sight-reading, remembering complicated scores, and improvisation became evident as he played, and altered, classical, jazz and popular music. He produced his own shows and versions of "The Mikado" and "Carmen," and performed as piano soloist with his school orchestra and the State Symphony Orchestra.

He reveled in music while excelling in athletics and the classical subjects taught at the 300-year-old Boston Latin School.

At Harvard University, Bernstein studied piano and composition, but developed a serious interest in composing only after meeting American composer Aaron Copland.

Bibliography

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from Medical Documents: A Survey. *Artificial Intelligence in Medicine*, *33*(2), 21. http://doi.org/10.1016/j.artmed.2004.07.017

Alguliyev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2015). An unsupervised approach to generating generic summaries of documents. *Applied Soft Computing*, *34*, 236–250. http://doi.org/10.1016/j.asoc.2015.04.050

Alonso, O. (2012). Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, 1–20. Retrieved from http://dx.doi.org/10.1007/s10791-012-9204-1

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *ACM SIGIR Forum*, *42*(2), 9. http://doi.org/10.1145/1480506.1480508

Ando, R. K., Boguraev, B. K., Byrd, R. J., & Neff, M. S. (2000). Multi-document summarization by visualizing topical content. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization* (pp. 79–98). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1567564.1567573

Atkinson, J., & Munoz, R. (2013). Rhetorics-based multi-document summarization. *Expert Systems with Applications*, *40*(11), 4346–4352. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417413000304

Baker, K. M. (1975). Condorcet, From Natural Philosophy to Social Mathematics. Retrieved from http://philpapers.org/rec/BAKCFN

Banko, M., Mittal, V. O., & Witbrock, M. J. (2000). Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 318–325). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1075218.1075259

Barry, C. L. (1994). User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=3986743

Bhaskar, P., & Bandyopadhyay, S. (2010). A Query FocuBhaskar, P., & Bandyopadhyay, S. (2010). A Query Focused Multi Document Automatic Summarization. In PACLIC (pp. 545–554).sed Multi Document Automatic Summarization. In *PACLIC* (pp. 545–554).

Bonnie Dorr, Christof Monz, Douglas Oard, Stacy President, David Zajic, R. S. (2004). Extrinsic Evaluation of Automatic Metrics for Summarization. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.2596

Bosma, W. E. (2005). Query-Based Summarization using Rhetorical Structure Theory. In T. van der Wouden, M. Po?, H. Reckman, & C. Cremers (Eds.), *15th Meeting of CLIN* (pp. 29–44). LOT.

Boydell, O., & Smyth, B. (2007). From social bookmarking to social summarization. In *Proceedings of the 12th international conference on Intelligent user interfaces - IUI '07* (p. 42). New York, New York, USA: ACM Press. http://doi.org/10.1145/1216295.1216311

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, *31*(5), 675–685. http://doi.org/http://dx.doi.org/10.1016/0306-4573(95)00052-I

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluation the Role of Bleu in Machine Translation Research. In *EACL* (Vol. 6, pp. 249–256).

Carbonell, J., & Goldstein, J. (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*. Melbourne, Australia: ACM. http://doi.org/10.1145/290941.291025

Carlson, L., Marcu, D., & Okurowski, M. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (Eds.), *Current and New Directions in Discourse and Dialogue SE - 5* (Vol. 22, pp. 85–112). Springer Netherlands. http://doi.org/10.1007/978-94-010-0019-2_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd editio). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden Markov models. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New Orleans, Louisiana, United States: ACM. http://doi.org/10.1145/383952.384042

Daumé, H., & Marcu, D. (2006). Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 305–312). Sydney, Australia: Association for Computational Linguistics. http://doi.org/10.3115/1220175.1220214

de Oliveira, P. C. F. (2005). How to Evaluate the" goodness" of Summaries Automatically. University of Surrey.

Diakopoulos, N., Choudhury, M. De, & Naaman, M. (2012). Finding and Assessing Social Media Information Sources in the Context of Journalism. *Business*, 2451–2460.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264–285.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, *22*, 457–479.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power

analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://doi.org/10.3758/BF03193146

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating named entities in Twitter data with crowdsourcing, 80–88. Retrieved from http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=1866696.1866709

Gao, W., Li, P., & Darwish, K. (2012). Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1173–1182). New York, NY, USA: ACM. http://doi.org/10.1145/2396761.2398417

Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008). Summarization System Evaluation Revisited: N-Gram Graphs. *ACM Transactions on Speech and Language Processing*, *5*(3), 1–39. http://doi.org/10.1145/1410358.1410359

Glaser, A., & Schütze, H. (2012). Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 276–285).

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19–25). New Orleans, Louisiana, United States: ACM. http://doi.org/10.1145/383952.383955

Goyal, P., Behera, L., & McGinnity, T. M. (2013). A Context-Based Word Indexing Model for Document Summarization. *Knowledge and Data Engineering, IEEE Transactions on*. http://doi.org/10.1109/TKDE.2012.114

Grady, C., & Lease, M. (2010). Crowdsourcing document relevance assessment with Mechanical Turk, 172–179. Retrieved from http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=1866696.1866723

Hachey, B. (2009). Multi-document summarisation using generic relation extraction. *Proceedings of*

the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. Singapore: Association for Computational Linguistics.

Harabagiu, S., Hickl, a, & Lacatusu, F. (2007). Satisfying information needs with multi-document summaries. *Information Processing & Management*, *43*(6), 1619–1642. http://doi.org/10.1016/j.ipm.2007.01.004

Harabagiu, S., & Lacatusu, F. (2002). Generating single and multi-document summaries with GISTexter. In *Document Understanding Conference 2002 (DUC'02)* (pp. 30–38).

Harman, D., Steinberger, J., Poesio, M., Kabadjov, M. A., & Ježek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing & Management*, *43*(6), 1663–1680. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306457307000428

Hirohata, M., Shinnaka, Y., Iwano, K., & Furui, S. (2005). Sentence extraction-based presentation summarization techniques and evaluation metrics. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)* (pp. 1065–1068).

Hosseini, M., Cox, I. J., Milić-Frayling, N., Kazai, G., & Vinay, V. (2012). On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, & F. Silvestri (Eds.), *Advances in Information Retrieval* (7224th ed., pp. 182–194). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-28997-2_16

Hovy, E., & Lin, C. Y. (1999). Automated Text Summarization in SUMMARIST. Advances in Automatic Text Summarization, 81-94. MIT Press.

Hu, M., Sun, A., & Lim, E.-P. (2007). Comments-Oriented Blog Summarization by Sentence Extraction. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM'07)*. Lisboa, Portugal.: ACM . http://doi.org/10.1145/1321440.1321571

Hu, M., Sun, A., & Lim, E.-P. P. (2008). Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st Annual International ACM Conference on Research and Development in Information Retrieval*

(SIGIR'08) (pp. 291–298). Singapore: ACM. http://doi.org/10.1145/1390334.1390385

Hu, P., Ji, D., Sun, C., Teng, C., & Zhang, Y. (2011a). Improving Document Summarization by Incorporating. *Language*, 499–508. http://doi.org/978-3-642-25630-1

Hu, P., Ji, D., Sun, C., Teng, C., & Zhang, Y. (2011b). Improving Document Summarization by Incorporating Social Contextual Information. In M. V. M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, & H. Khelalfa (Eds.), *7th Asia Information Retrieval Societies Conference (AIRS 2011)* (pp. 499–508). Dubai, United Arab Emirates: Springer.

Hu, P., Sun, C., Wu, L., Ji, D.-H., & Teng, C. (2011). Social Summarization via Automatically Discovered Social Context. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 483–490). Chiang Mai, Thailand.

Inouye, D., & Kalita, J. K. (2011). Comparing Twitter Summarization Algorithms for Multiple Post Summaries. *Privacy, Security, Risk and Trust (Passat), 2011 Ieee Third International Conference on and 2011 Ieee Third International Conference on Social Computing (Socialcom)*. http://doi.org/10.1109/PASSAT/SocialCom.2011.31

Jagadeesh, J., Pingali, P., & Varma, V. (2007). Capturing sentence prior for query-based multi-document summarization, 798–809. Retrieved from http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=1931390.1931465

Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization* (pp. 51–59).

Jones, N., Brun, A., & Boyer, A. (2011). Comparisons Instead of Ratings: Towards More Stable Preferences. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 451–456). IEEE. http://doi.org/10.1109/WI-IAT.2011.13

Kan, M.-Y., & Klavans, J. L. (2002). Using librarian techniques in automatic text summarization for information retrieval. *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. Portland, Oregon, USA: ACM. http://doi.org/10.1145/544220.544227

Keikha, M., Park, J. H., & Croft, W. B. (2014). Evaluating answer passages using summarization

measures. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14* (pp. 963–966). New York, New York, USA: ACM Press. http://doi.org/10.1145/2600428.2609485

Kim, S., Oh, J. S., & Oh, S. (2008). Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–15. http://doi.org/10.1002/meet.1450440256

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 453). New York, New York, USA: ACM Press. http://doi.org/10.1145/1357054.1357127

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 207). New York, New York, USA: ACM Press. http://doi.org/10.1145/2470654.2470684

Kothari, A., Magdy, W., Darwish, K., Mourad, A., & Taei, A. (2013). Detecting Comments on News Articles in Microblogs. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Kozorovitzky, A., & Kurland, O. (2009). From "Identical" to "Similar": Fusing Retrieved Lists Based on Inter-document Similarities. In L. Azzopardi, G. Kazai, S. Robertson, S. Rüger, M. Shokouhi, D. Song, & E. Yilmaz (Eds.), *Advances in Information Retrieval Theory SE - 19* (Vol. 5766, pp. 212–223). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-04417-5_19

Kruengkrai, C., & Jaruskulchai, C. (2003). Generic text summarization using local and global properties of sentences. *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. http://doi.org/10.1109/WI.2003.1241194

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. http://doi.org/10.1016/j.asoc.2014.03.041

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, Washington, United States: ACM. http://doi.org/10.1145/215206.215333

Kushniruk, A. W., Kan, M.-Y., McKeown, K., Klavans, J., Jordan, D., LaFlamme, M., & Patel, V. L. (2002). Usability evaluation of an experimental text summarization system and three search engines: implications for the reengineering of health care interfaces. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 420–4. Retrieved from /pmc/articles/PMC2244493/?report=abstract

Leuski, A., Lin, C.-Y., & Hovy, E. (2003). iNeATS: interactive multi-document summarization. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2*. Sapporo, Japan: Association for Computational Linguistics. http://doi.org/10.3115/1075178.1075197

Li, X., Du, L., & Shen, Y.-D. (2013). Update Summarization via Graph-Based Sentence Ranking. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 1162–1174. http://doi.org/10.1109/TKDE.2012.42

Lin, C.-Y. (2004a). Looking for a few good metrics: ROUGE and its evaluation. In *NTCIR Workshop*.

Lin, C.-Y. (2004b). Rouge: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)* (pp. 74–81). Barcelona, Spain.

Lin, C.-Y., & Hovy, E. (1997). Identifying topics by position. *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, DC: Association for Computational Linguistics. http://doi.org/10.3115/974557.974599

Lin, C.-Y., & Hovy, E. (2002). Automated multi-document summarization in NeATS. (M. Marcus, Ed.), *Proceedings of the Second International Conference on Human Language Technology Research*. San Diego, California: Morgan Kaufmann Publishers Inc.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence

statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (pp. 71–78). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1073445.1073465

Liu, F., & Liu, Y. (2010). Exploring correlation between ROUGE and human evaluation on meeting summaries. *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(1), 187–196.

Lloret, E., & Palomar, M. (2010, January 1). Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *International Journal of Informatica*. Retrieved from http://repository.dlsi.ua.es/307/1/title-acks.pdf.pdf

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, *37*(1), 1–41. http://doi.org/10.1007/s10462-011-9216-z

Lloret, E., Plaza, L., & Aker, A. (2013). Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, *47*(2), 337–369. http://doi.org/10.1007/s10579-012-9198-8

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. http://doi.org/10.1147/rd.22.0159

Lv, Y., & Zhai, C. (2011). When documents are very long, BM25 fails! In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 1103–1104). New York, NY, USA: ACM. http://doi.org/10.1145/2009916.2010070

Mackie, S., McCreadie, R., Macdonald, C., & Ounis, I. (2014). On choosing an effective automatic evaluation metric for microblog summarisation. In *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14* (pp. 115–124). New York, New York, USA: ACM Press. http://doi.org/10.1145/2637002.2637017

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53*(5), 327–342.

http://doi.org/10.1002/asi.10049

Mani, I. (2001). *Automatic summarization*. (R. Mitkov, Ed.) (Vol. 3). John Benjamins Publishing.

Mani, I., & Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence* (pp. 622–628). AAAI Press.

Mani, I., & Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *American Association for Artificial Intelligence (AAAI)* (pp. 821–826).

Mani, I., & Bloedorn, E. (1999). Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, *1*(1–2), 35–67. http://doi.org/10.1023/a:1009930203452

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., & Sundheim, B. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics* (pp. 77–85). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/977035.977047

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Interdisciplinary Journal for the Study of Discourse Text*, *8*(3). http://doi.org/10.1515/text.1.1988.8.3.243

Marcu, D. (1997a). From discourse structures to text summaries. In *ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (Vol. 97, pp. 82–88). Madrid, Spain.

Marcu, D. (1997b). The rhetorical parsing of natural language texts. *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics. http://doi.org/10.3115/979617.979630

McKeown, K. R., Chang, S.-F., Cimino, J., Feiner, S., Friedman, C., Gravano, L., … Teufel, S. (2001). PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*. Roanoke, Virginia, United States: ACM. http://doi.org/10.1145/379437.379722

McKeown, K., & Radev, D. R. (1995). Generating summaries of multiple news articles. In

Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 74–82). New York, NY, USA: ACM. http://doi.org/10.1145/215206.215334

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics. http://doi.org/10.3115/1219044.1219064

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP* (Vol. 4). Barcelona, Spain.

Mithun, S. (2010). Exploiting rhetorical relations in blog summarization. *Advances in Artificial Intelligence*, 388–392.

Mithun, S., & Kosseim, L. (2009). Summarizing blog entries versus news texts. *Proceedings of the Workshop on Events in Emerging Text Types*. Borovets, Bulgaria: Association for Computational Linguistics.

Mollá, D. (2010). A corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Workshop* (Vol. 8, pp. 76–80).

Morris, A. H., Kasper, G. M., & Adams, D. A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, *3*(1), 17–35.

Myaeng, S. H., & Jang, D.-H. (1999). Development and evaluation of a statistically-based document summarization system. *Advances in Automatic Text Summarization*, 61–70.

Nenkova, A., & Louis, A. (2008). Can You Summarize This? Identifying Correlates of Input Difficulty for Generic Multi-Document Summarization. In *46th Annual Meeting of the Association for Computational Linguistics*.

Nenkova, A., & McKeown, K. R. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval*, *5*(2–3), 130. http://doi.org/10.1561/1500000015

Nenkova, A., & Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid

Method. Retrieved from http://academiccommons.columbia.edu/catalog/ac:161762

Nenkova, A., Passonneau, R., & McKeown, K. (2007). The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Trans. Speech Lang. Process.*, *4*(2). http://doi.org/10.1145/1233912.1233913

Nichols, J., Mahmud, J., & Drews, C. (2012). Summarizing Sporting Events Using Twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (pp. 189–198). New York, NY, USA: ACM. http://doi.org/10.1145/2166966.2166999

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing. In *Proceedings of the international conference on Multimedia information retrieval - MIR '10* (p. 557). New York, New York, USA: ACM Press. http://doi.org/10.1145/1743384.1743478

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: bringing order to the web.* Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/

Palshikar, G., Deshpande, S., & Athiappan, G. (2012). Combining Summaries Using Unsupervised Rank Aggregation. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing SE - 32* (Vol. 7182, pp. 378–389). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-28601-8_32

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (p. 311). Morristown, NJ, USA: Association for Computational Linguistics. http://doi.org/10.3115/1073083.1073135

Parapar, J., López-Castro, J., & Barreiro, Á. (2010). Blog snippets: a comments-biased approach. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 711–712). New York, NY, USA: ACM. http://doi.org/10.1145/1835449.1835578

Park, J., Fukuhara, T., Ohmukai, I., Takeda, H., & Lee, S. (2008). Web content summarization using social bookmarks: a new approach for social summarization. In *Proceedings of the 10th ACM*

workshop on Web information and data management (pp. 103–110). New York, NY, USA: ACM. http://doi.org/10.1145/1458502.1458519

Park, S., Mohammadi, G., Artstein, R., & Morency, L.-P. (2012). Crowdsourcing micro-level multimedia annotations. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia - CrowdMM '12* (p. 29). New York, New York, USA: ACM Press. http://doi.org/10.1145/2390803.2390816

Ping, C., & Verma, R. (2006). A Query-Based Medical Information Summarization System Using Ontology Knowledge. In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on* (pp. 37–42).

Plaza, L., Díaz, A., & Gervás, P. (2011). A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, *53*, 1–14. http://doi.org/10.1016/j.artmed.2011.06.005

Pollock, J. J., & Zamora, A. (1975). Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, *15*(4), 226–232.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *14*(3), 130–137.

Qiu, L.-Q., & Pang, B. (2008). Analysis of Automated Evaluation for Multi-document Summarization Using Content-Based Similarity. *Digital Society, 2008 Second International Conference on the*. http://doi.org/10.1109/ICDS.2008.9

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Comput. Linguist.*, *28*(4), 399–408. http://doi.org/10.1162/089120102762671927

Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing Management*, *40*(6), 919–938. http://doi.org/10.1016/j.ipm.2003.10.006

Radev, D. R., & McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Comput. Linguist.*, *24*(3), 470–500.

Rath, G. J., Resnick, A., & Savage, T. R. (1961). The formation of abstracts by the selection of sentences. *American Documentation*, *12*(2), 139–141. http://doi.org/10.1002/asi.5090120210

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North America Chapter of the ACL.*

Robertson, S. ., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, *36*(1), 95–108. http://doi.org/10.1016/S0306-4573(99)00046-1

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, *33*(2), 193–207. http://doi.org/http://dx.doi.org/10.1016/S0306-4573(96)00062-3

Sanderson, M., Paramita, M. L., Clough, P., & Kanoulas, E. (2010). Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 555–562). New York, NY, USA: ACM. http://doi.org/10.1145/1835449.1835542

Saracevic, T. (2007). Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933. http://doi.org/10.1002/asi.20682

Savolainen, R., & Kari, J. (2013). User-defined relevance criteria in web searching. *Journal of Documentation*, *62*(6), 685–707. Retrieved from http://www.emeraldinsight.com.ezproxy.lib.rmit.edu.au/doi/full/10.1108/00220410610714921

Sharifi, B., Hutton, M. A., & Kalita, J. (2010). Automatic Summarization of Twitter Topics. In *National Workshop on Design and Analysis of Algorithms, Tezpur, India* (pp. 121–128).

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? attributing scientific work to citations. Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.3178

Silber, H. G., & McCoy, K. F. (2000). Efficient text summarization using lexical chains. *Proceedings of the 5th International Conference on Intelligent User Interfaces*. New Orleans, Louisiana, United States: ACM. http://doi.org/10.1145/325737.325861

Sizov, G. (2010). *Extraction-Based Automatic Summarization*. *Science And Technology*. Norwegian University of Science and Technology.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks, 254–263. Retrieved from http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=1613715.1613751

Soe-Tsyr, Y., & Jerry, S. (2005). Ontology-based structured cosine similarity in document summarization: with applications to mobile audio-based knowledge management. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *35*(5), 1028–1040.

Song, X., Chi, Y., Hino, K., & Tseng, B. L. (2007). Summarization System by Identifying Influential Blogs. *ICWSM 2007*. Boulder, Colorado, U.S.A.

Song, X., Cohn, T., & Specia, L. (2013). BLEU deconstructed: Designing a better MT evaluation metric. *International Journal of Computational Linguistics and Applications*, *4*(2), 29.

Spärck Jones, K. (1993). What might be in a summary. In G. Knorz, J. Krause, & C. Womser-Hacker (Eds.), *Information Retrieval 93: From Modeling to Application* (pp. 9–26). University of Konstanz Verlag.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, *43*(6), 1449–1481. http://doi.org/10.1016/j.ipm.2007.03.009

Štajner, T., Thomee, B., Popescu, A.-M., Pennacchiotti, M., Jaimes, A., Stajner, T., … Jaimes, A. (2013). Automatic selection of social media responses to news. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 50–58). New York, NY, USA: ACM. http://doi.org/10.1145/2487575.2487659

Steinberger, J., & Ježek, K. (2012). Evaluation measures for text summarization. *Computing and*

Informatics, *28*(2), 251–275.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis* (Vol. 2, pp. 2–6).

Svore, K. M., & Burges, C. J. C. (2009). A machine learning approach for improved BM25 retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (p. 1811). New York, New York, USA: ACM Press. http://doi.org/10.1145/1645953.1646237

Tao, T., Wang, X., Mei, Q., & Zhai, C. (2006). Language model information retrieval with document expansion. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. New York, New York: Association for Computational Linguistics. http://doi.org/10.3115/1220835.1220887

Teufel, S. (2000). Argumentative Zoning: Information extraction from scientific text. Citeseer.

Teufel, S. (2001). Task-based evaluation of summary quality: Describing relationships between scientific papers. In *In Workshop Automatic Summarization, NAACL*. Citeseer.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist.*, *28*(4), 409–445. http://doi.org/10.1162/089120102762671936

Thakkar, K. S., Dharaskar, R. V, & Chandak, M. B. (2010). Graph-Based Algorithms for Text Summarization. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology* (pp. 516–519). IEEE. http://doi.org/10.1109/ICETET.2010.104

Thomas, D. R. (2006, June). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*. http://doi.org/10.1177/1098214005283748

Tideman, T. N. (1987). Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, *4*(3), 185–206. http://doi.org/10.1007/BF00433944

Varadarajan, R., & Hristidis, V. (2005). Structure-based query-specific document summarization. In

Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 231–232). New York, NY, USA: ACM. http://doi.org/10.1145/1099554.1099602

Varadarajan, R., & Hristidis, V. (2006). A system for query-specific document summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 622–631). New York, NY, USA: ACM. http://doi.org/10.1145/1183614.1183703

Viera, A. J., & Garret, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, *37*(5), 360–363. Retrieved from http://virtualhost.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf

Volkovs, M. N., Larochelle, H., & Zemel, R. S. (2012). Learning to rank by aggregating expert preferences. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12* (p. 843). New York, New York, USA: ACM Press. http://doi.org/10.1145/2396761.2396868

Volkovs, M. N., & Zemel, R. S. (2014). New learning methods for supervised and unsupervised preference aggregation. *The Journal of Machine Learning Research*, *15*(1), 1135–1176. Retrieved from http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=2627435.2638572

Wan, X., & Xiao, J. (2010). Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *ACM Transaction of Information Systems*, *28*(2), 1–34. http://doi.org/10.1145/1740592.1740596

Wan, X., & Yang, J. (2006). Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 181–184). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1614049.1614095

Wan, X., Yang, J., & Xiao, J. (2007). CollabSum: exploiting multiple document clustering for

collaborative single document summarizations. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, The Netherlands: ACM. http://doi.org/10.1145/1277741.1277768

Wang, D., Zhu, S., Li, T., Chi, Y., & Gong, Y. (2011). Integrating Document Clustering and Multidocument Summarization. *ACM Transaction Knowledge and Discovery Data*, *5*(3 (14)), 26 pages. http://doi.org/10.1145/1993077.1993078

Webber, W., Moffat, A., & Zobel, J. (2008). Statistical power in retrieval experimentation. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08* (p. 571). New York, New York, USA: ACM Press. http://doi.org/10.1145/1458082.1458158

Wei, Z., Gao, W., El-Ganainy, T., Magdy, W., & Wong, K.-F. (2014). Ranking model selection and fusion for effective microblog search. In *Proceedings of the first international workshop on Social media retrieval and analysis - SoMeRA '14* (pp. 21–26). New York, New York, USA: ACM Press. http://doi.org/10.1145/2632188.2632202

Witbrock, M. J., & Mittal, V. O. (1999). Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 315–316). New York, NY, USA: ACM. http://doi.org/10.1145/312624.312748

Wolf, F., & Gibson, E. (2004). Paragraph-, word-, and coherence-based approaches to sentence ranking: a comparison of algorithm and human performance. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1218955.1219004

Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., & Li, J. (2011). Social Context Summarization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 255–264). New York, NY, USA: ACM. http://doi.org/10.1145/2009916.2009954

Yoo, I., Hu, X., & Song, I.-Y. (2006). Integrating biomedical literature clustering and summarization

approaches using biomedical ontology. *Proceedings of the 1st International Workshop on Text Mining in Bioinformatics*. Arlington, Virginia, USA: ACM. http://doi.org/10.1145/1183535.1183545

Yulianti, E. (2013). *Tweet-Biased Summarization*. Master Dissertation, School of Computer Science and Information Technology, RMIT University, Victoria, Australia.

Yulianti, E., Huspi, S., & Sanderson, M. (2016). Tweet-biased summarization. *Journal of the Association for Information Science and Technology*, *67*(6), 1289–1300. http://doi.org/10.1002/asi.23496

Zha, H. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 113–120). Tampere, Finland: ACM. http://doi.org/10.1145/564376.564398

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W.-Y. (2005). Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 504–511). ACM.