Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Dissertation Title

John Carbeck
16309095

April 28, 2020

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
BA (Computer Science)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`. I

have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at

`http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ______________________ Date: ______________________

Abstract

A short summary of the problem investigated, the approach taken and the key findings. This should be around 400 words, or less.

This should be on a separate page.

Acknowledgements

Thanks Mum!

You should acknowledge any help that you have received (for example from technical staff), or input provided by, for example, a company.

Contents

List of Figures

List of Tables

Nomenclature

A	Area of the wing	m^2
B		
C	Roman letters first, with capitals…	
a	then lower case.	
b		
c		
Γ	Followed by Greek capitals…	
α	then lower case greek symbols.	
β		
ϵ		
TLA	Finally, three letter acronyms and other abbreviations arranged alphabetically	

If a parameter has a typical unit that is used throughout your report, then it should be included here on the right hand side.

If you have a very mathematical report, then you may wish to divide the nomenclature list into functions and variables, and then sub- and super-scripts.

Note that Roman mathematical symbols are typically in a serif font in italics.

1 Introduction

1.1 Motivation

Over half of the world has access to the internet, and every one of those internet users faces information overload. Information overload is defined as the difficulty in dealing with an information load of great quantity, complexity, redundancy, contradiction and inconsistency (Gross 1964, Roetzel 2019). Ever since humans have created information, there have been systems developed to handle the storage and retrieval of that information. With modern information and communication technologies the amount of information is exploding and so is the problem of information overload. Approaches to remedy information overload aim to reduce the amount of incoming information to a recipient as well as enhance the recipient's information processing capabilities (Soucek & Moser 2010). Most of the content available on the internet is unstructured such as: images, videos, and bodies of text. Thus systems that are developed to help reduce information overload must therefore be able to represent the content of unstructured information in order to determine relevance to limit information and to enhance user processing capabilities.

Information retrieval (IR) systems attempt to retrieve unstructured information from a large collection, based on an expressed information need of a user. Most commonly users express their information needs via queries. IR relies on two tasks that impact the effectiveness of the retrieval systems. Inorder for information to be retrieved the information must be categorised and represented based on the content they contain. The degree of comprehension of the content in the representation affects the system's ability to identify relevant content (Chiaramella 2000). The user must also be able to express their information needs in comprehensible form to the IR system, or the system must be able to extrapolate information needed from natural language processing and query context (Carpineto & Romano 2012). One of the most visible forms of IR systems are search engines like Google. Search engines attempt to perform IR on large volumes of content that exists on the internet, but even well expressed and specific queries produce millions of relevant results, reducing content presented to the users but not significantly mitigating information overload. Thus IR systems have been developed to reduce content further beyond list of relevant material and have been implemented to serve specific domains performing retrieval on smaller sets of information.

Domain specific IR systems aim to improve the form of the retrieved information as well as relevance of documents to the specific information needs of the user. Within a specific domain this can be done with use of domain specific models that represent topic and terms and their relations. Domain models take the form of ontologies or knowledge bases. Domain models are used in IR systems to expand queries from semantic term relations, extract topics of content, and can be used to form abstract representations of retrieved content in the form of abractive summaries.

Automatic text summarisation is a form of IR that creates a summary from one or many documents, maintaining key information from the original text(s). Automatic text summarisation is an effective method for reducing information overload as summaries represent relevant information in a digestible condensed form. Text summarisation can also be query based, allowing for users to request a summary based on an information need. Query text summarisation can be further enhanced by models of user's knowledge or interest in a domain. Using user models, personalised summaries are produced with regard to both a user query and the user's knowledge or interest. Not only does this form of IR further reduce information presented to the user based on what is most salient to them, but it can also lead a user towards information that is novel or of greater interest. Personalised query based summarisation systems also allow for more effective feedback loops then non personalised query based IR. The system can present with its summarisation its interpretation of the user interest or knowledge, facilitating interactive information retrieval (Chiaramella 2000). The interpretation of knowledge can then be adjusted by the user to provide a summary that better serves the users' information needs.

Generalised text summarisation methods coverage of key content suffers from the complexity of forming a summary of short length from a large set of documents (Goldstein et al. 2000) et al., 2000). Generalised text summarisation methods also fail to reflect term importance and relations when applied to documents in a specific domain. Thus domain specific text summarisation methods attempt to provide better concept extraction, document representation, and summary formation from the use of domain models, such as ontologies or semantic based knowledge bases. Domain specific text summarisation has been applied to domains such as medicine (Sarker et al. 2013) and law (Galgani et al. 2012), but these domain models are hand crafted or supervised by experts from those fields. Domain specific summarisation is limited based on the model's comprehension of the domain. General models such as knowledge bases (supervised models created from semantic relationships) fail to comprehend domain specific term relationships and term significance. Thus common-sense knowledge bases are unable to replace expert crafted ontologies. Systems that provide domain specific personalised automatic text summarisation also use domain models in the creation of user interest or knowledge models (Ge et al. 2012), these systems are twice as reliant on domain models and their coherency. The majority of content online does not relate to domains with supplied or are easily generated ontologies.

For example, Wikipedia offers over 6 million articles in English. Due to the plethora of relevant and related documents for a given topic in a large domain, as well as the cyclic relations between documents, information overload and disorientation are common problems experienced by users researching a particular topic. World War II as an example contains 26,388 directly related pages. Even within a small domain such as the Watergate Scandal, there are 32 relevant pages. Some pages provide a general overview of a topic, but when reading or learning on a sub-topic of a broad topic, many of the pages that are related lay latent. These topics also lack formal ontologies so methods that reduce information overload from domain specific summarisation cannot be applied to them.

Therefore there is a need for domain specific text summarisation systems that are unsupervised. This would allow for the benefits of personalised domain specific summarisation to occur on the majority of content online, that do not have formalised domain models.

1.2 Research Question

The main problem with the application of domain specific personalised summarisation much of the domains of content on the internet is that existing methods are reliant on supervised domain models. While these existing methods have proven effective on domains with formal ontologies, the majority of available online content lacks formal ontologies. Work has been done on the automatic formation of ontologies (Bedini & Nguyen 2007), but many of the state of the art techniques are reliant on curated domain corpuses, and others require expert validation. Automatically generated ontologies still fail to achieve the quality of domain expert generated ones. Therefore existing methods that work with expert generated ontologies, can only perform worse when used with automatically generated ones.

The tasks in automatically generating ontologies are similar to the formation of topic representations used in extractive text summarisation methods. Extractive summarisation methods create topic representation from: the extraction of terms, creation of topics based on semantic relationships or frequency of terms, and weighting relations of terms to topics. These tasks are very similar to those done in automatic ontology generation such as: extraction of concepts attributes and relations from a corpus source and analysis of extracted content to determine relations between content or ontologies, as described in (Bedini & Nguyen 2007). These similarities suggest that existing extractive summarisation methods that use topic representations may lend themselves to performing domain specific personalised summarisation. This project explores this by addressing the following question:

To what extent can existing automatic extractive summarization methods be used to provide domain specific personalised summaries, independent of domain specific ontologies and semantic models?

The extent to which existing automatic extractive methods can be used is reliant on how well the proposed system can match the performance of state-of-the-art extractive summarisation methods. The extent is also reliant on systems ability to produce personalise summaries based on an given individual user knowledge model within a specific domain.

1.3 Objectives

The research question can be broken down into three objectives, which when completed, produced an unsupervised personalised domain independent summarisation system from existing systems and a determination of the efficacy of such a system. The three objectives are: to complete a review of the classifications, tasks, and methods of automatic summarisation; design and implement a system which performs domain specific personalised summaries without ontologies; and evaluate the extent to which the designed system is effective. These objectives together result in the answering of the research question. These three objectives are presented in detail below.

O1: Review of Automatic Text Summarisation

The first objective is to perform a review of automatic text summarisation in order identify the classifications, methods, and approaches to tasks of summarisation. This review will be used in O2, to design a system that fulfills the requirements of the research question. To complete this objective system a broad review of the field of automatic text summarisation must be conducted. The aim of this review is to consider all approaches that may fulfill the requirements of the system. The review also serves to inform the approach of evaluating the proposed summarisation system. Thus the objectives of the literature view are:

- Determine the classification of summarisation approaches.
- Identify the tasks involved in summarisation.
- Review of extractive summarisation methods.
- Review of evaluation methods for summarisation systems.

O2: Design and Implement a Summarization System

To answer the research question a system must be designed and implemented which performs domain specific personalised summaries without ontologies. First the requirements, classifications and tasks need to be defined to outline a system which will achieve this objective. Existing methods of summarisation can then be identified from review of automatic text summarisation, to construct a summarisation system design from existing methods. The design must then be implemented, using the selected methods from their respective explanations. The objectives of the design and implementation objective for this project are:

- the classification, and requirements needed by a system to perform domain specific personalised summarisation without formal ontologies
- Use existing methods of summarisation to identify a set of tasks that fit the classifications of the system and fulfill the requirements of the system.
- Select methods for the identified tasks from existing or a combination of existing approaches of automatic text summarisation to create a system design.
- From the design implement the selected methods, from their mathematical formalisation, algorithms or from libraries that contain the functionality, to produces a personalised summaries independent of domain models

O3: Evaluation of Proposed System

The final objective of this project is to evaluate the extent to which the proposed system can perform personalised summaries in a given domain without the use of formal ontologies. This can be described by the following objectives:

- Compare the performance of the proposed system to other state of the art systems in order to efficacy of the implementation and design.

- Examine the system's ability to personalise summaries based on specific model states to determine further work, and successful components of the system

1.4 Overview

This chapter has articulated the motivation, research question, and objectives of this project. The following chapters will build on this chapters foundation, further explaining the background, reasoning, and approach to determining the extent that existing extractive summarisation method can be used to perform specific domain personalised text summarisation without the need for supervised domain models. The system that was designed and implemented for this project, achieves competitive performance to other extractive systems and is found to be viable for use as a domain specific personalised summarisation system, independent of supervised domain models. The process to which the design of the proposed system was informed, developed, constructed and evaluated is presented in following chapters of this paper.

Chapter 2 presents the taxonomy, tasks, and approaches to automatic text summarisation. This chapter provides the necessary background for the reasoning used in the construction of the system design and for the methods used to implement the final system. The information reviewed in this chapter was used in informing the analysis and method for the design, implementation, and evaluation chapters.

Chapter 3 describes the methodology used in creating the design of the needed system. The classification and requirements of the system were defined inorder to identify the necessary tasks that the system needed to perform. Then methods from existing systems were selected to perform the identified tasks, based on the system requirements, compatibility with other methods, and performance. This chapter produces a system design that was used in the following implementation chapter.

Chapter 4 outlines the implementation of the proposed system from existing methods. The summarisation system was implemented from the process described in each method's respective literature. Some methods' implementation was done via the use of libraries, while other methods required bespoke implementation based on the mathematical or pseudocode presented in the papers that present them. The implementation produced is a set of python classes which encapsulate selected methods for the system. The classes are used together to provide a system that performs personalised extractive summarisation on domains without formal ontologies.

Chapter 5 presents the two methods used to determine the extent that this system can provide personalised summarisation on a domain independent of domain specific models. From the comparative evaluation of this system summarisation performance with state of the art extractive system, the system presents a competitive method of performing extractive summarisation. From the examination of the system on a specific domain set of document, the system is shown to produce interprobable personalised summaries from a context free and contextual queries, demonstrating that this system can be used to perform domain specific personalised summarisation on a specific domain.

Chapter 6 presents the limitation of the system and opportunities for future work for the design and evaluation.

Chapter 7 reviews the material presentend in this paper as well as the objectives set out for this project and the extent to which they were met.

2 Background

2.1 Automatic Text Summarisation

Automatic text summarisation can be approached in many different ways. Generally the aim is to produce a summary, defined as "a text that is produced from one or more texts, that conveys important information in the original text(s)"(Radef et al. 2002). Allahyari et al. (2017) define Automatic text summarisation as "the task of producing a concise and fluent summary while reserving key information content and overall meaning". Automatic text summarisation has many forms as each summarisation task uses different types of source documents, representation of content, and reasoning in producing a summary. The many forms of automatic text summarisation are discussed in this section.

2.1.1 Summary Characteristics

The context of the summarisation task must be addressed in order to best perform automatic text summarisation. Spärck-Jones (1999) in her taxonomy writes: "It is important to recognize the role of context factors because the idea of a general-purpose summary is manifestly an ignis factus". The three context factors she identifies are input factors, purpose factors, and output factors. Input factors are classification of the representation of input document(s) in terms of structure, genre, format, and unit. The purpose factors are the relationship between the source and the output of summarisation and are described as dealing with situation, audience, and use. The output factors define the form of output of summary and are largely driven by the input and purpose factors of the system.

2.1.2 Types of Summaries

Gambhir and Gupta (2017) as well as Orăsan (2019) present a recent taxonomy to classify types of summaries. These classifications are important to consider when selecting an existing summarization method for a specific task or when creating a new automatic summarisation method.

Single document and Multi-document Summarisation

The input to a summarisation system can either be a single document or a set of multiple documents. Single document summarization addresses the content of a single document and produces a summary of that single document. Multi-document summarization considers content from multiple documents and produces a summary of the discussed topics across all given documents.

Many of the techniques of single document summarisation can be used in multi-document summarisation. Goldstein et al. (2000) identifies that: the redundancy of information of topically-related documents is much greater than in a single document making anti-redundancy methods crucial, the compression ratio (i.e. summary length with respect to document set length) is much smaller adding more difficulty to summarisation as compression demand increases, as well as the increased amount of co-referencing in a set of multiple documents than single documents. Many recent approaches attempt to deal with these issues. Methods to handle these issues will be discussed in the Redundancy Reduction section in this chapter.

Extractive and Abstractive Summarisation

The output of an automatic summarisation is either extractive or abstractive. An extractive summary is created from a subset of sentences from the source document(s). The sentences selected are those that the summarisation method finds most salient in the original text, using a similarity or centrality metric. An abstractive summary uses semantic models to generate a new piece of text that covers the themes, concepts or terms of the examined document or documents. Abstractive summarization requires natural language processing to extract concepts from the source material and to create an abstract summary from concept and word semantic relationships. Extractive summarization is simpler than abstractive but is limited because not all information in a sentence might not relate for a summary.

Generic and Query-focused Summaries

The purpose of summarisation is either generic or query-focused. Generic summaries attempt to summarize the content of all the material in the document or documents. This is the most common form of summary and is often used with single document summarisation. Query-focused, also referred to as topic-focused or user-focused, provides a summarisation based on a described need. These are commonly used with multi-document summarisation as multiple documents often contain a variety of topics. In this form of automatic text summarization a query is used both for the retrieval of documents as well as for the generation of the summary.

Personalised summaries are a type of user-focused summary. Personalised summaries aim to produce a tailored summary based on a model of the user. Díaz and Gervás (2007) personalised summaries of newswire texts using a model of user interests based on keywords, domain-specific factors and user feedback. Li, Liu and Zhao (2015) suggest an update summarisation system which considers the novelty of the sentence by adding novelty as a variable to traditional integer linear programming approaches of summarisation.

Supervised and Unsupervised Automatic Summarization

Another distinction of summarization methods is supervised and unsupervised methods. Supervised methods require training from a pre-labeled data set. Supervised methods use two class classification algorithms, trained on labeled data, for selection of important content in source documents. Unsupervised methods are able to generate summaries from only using the source documents, and can therefore operate on new documents without the need for training. Unsupervised summarisation identifies relevant sentences using a set of heuristic rules and uses those sentences to generate a summary.

2.2 Tasks of Summarisation

2.2.1 Intermediate Representation

2.2.2 Sentence Scoring

2.2.3 Sentence Selection

2.3 Summarisation Methods

2.3.1 Topic Representation

2.3.2 Indicator Representation and Machine Learning

2.3.3 Comparison of Methods

2.4 Query focused Summarisation

2.5 Figures

Graphs, pictures and other images should be included in your report as a numbered, captioned figure. An example is given in Figure 2.1.

The figure and caption should be centred. The figure numbering starts at 1 at the beginning of each chapter. The caption should provide a brief description of what is being shown. The figure should appear in the document after it is referred to in the text. No figure should be included which is not referred to in the text. Ensure that the size and resolution of images imported from software are sufficient to read any text.

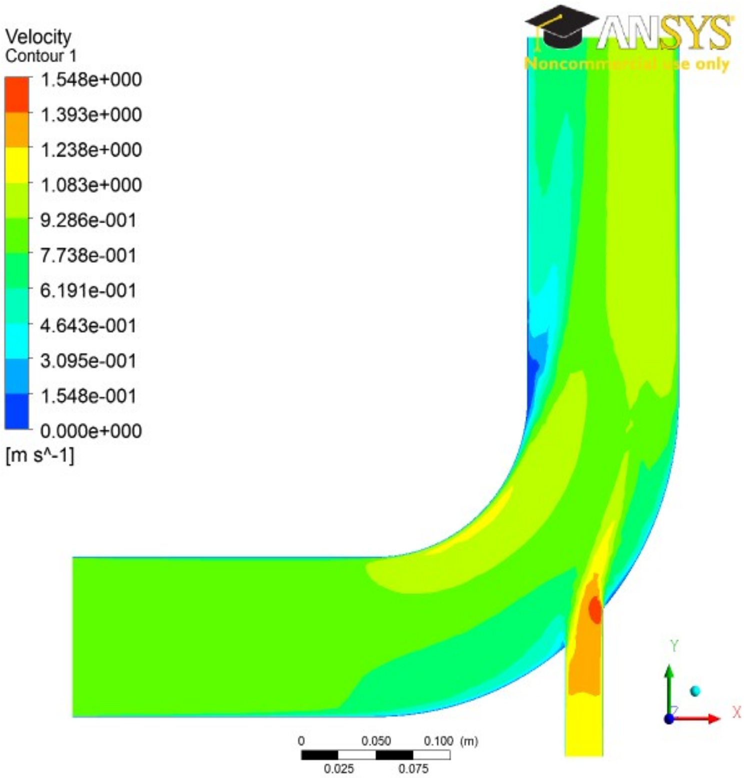

Figure 2.1: Velocity distribution on the mid-plane for an inlet velocity for case 1.

2.6 Tables

Tables are an important way of displaying your results; Table 2.1 is a sample table, adapted from the Master/Doctoral Thesis template at `http://www.latextemplates.com/cat/theses`, which was generated with this code:

```
\begin{table}[b]
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\textbf{Groups} & \textbf{Treatment X} & \textbf{Treatment Y} \\\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

Tables are numbered in the same way as figures. Typically tables also have a short caption, but this is not universally true. The number and caption appear above the table, not below as with figures. Again, no table should appear in the report which has not been referred to in the text. Tables should come after they are discussed in the text. The exact formatting of the table depends somewhat on the content of the table, but in general, the text in the table should be the same font and size as the main text.

2.7 Equations

All equations should be numbered sequentially. Do not restart the numbering at the beginning of each chapter. Unlike figures and tables, you may not need to refer to every equation in the text. You should take care to format equations properly. Do no simply try to use plain text. Use the equation layout facilities. An example of how equations should appear is shown in Equation 1. Here is the code for it:

```
\begin{equation}
\textrm{div}(\underline{u}) = \frac{\delta u}{\delta x} + \frac{\delta v}{\delta y} +
        \frac{\delta w}{\delta z} = 0
\label{sampleequation}
\end{equation}
```

$$\textrm{div}(\underline{u}) = \frac{\delta u}{\delta x} + \frac{\delta v}{\delta y} + \frac{\delta w}{\delta z} = 0 \quad (1)$$

2.8 Referencing published work

It is important to give appropriate credit to other people for the work that they have shared through publications. In fact, you must sign a declaration in your report stating that you understand the nature of plagiarism. As well as avoiding plagiarism,

Table 2.1: The effects of treatments X and Y on the four groups studied.

Groups	**Treatment X**	**Treatment Y**
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

citing results or data from the literature can strengthen your argument, provide a favourable comparison for your results, or even demonstrate how superior your work is.

There are many styles to reference published work. For example, the parenthetical style (which is also called the Harvard style) uses the author and date of publication (e.g. "Smith and Jones, 2001"). There is also the Vancouver (or the citation sequence) style, which is shown in this document. In the Vancouver style, the publications are cited using a bracket number which refers to the list in the References section at the end of the report. The references are listed in order that they are cited in the report. A variant is name sequence style in which the publications are referenced by number, but the list is arranged alphabetically. For example, the text might say: several studies have examined the sound field around tandem cylinders generated by flow**??**, while other investigations have focused on the effect of an applied sound field on the flow**?**. Papers from conference proceedings**?**, books**?** and technical reports**??** can be dealt with in the same style.

The Vancouver style has the advantage that it is a little more compact in the text and does not distract from the flow of the sentence if there are a lot of citations. However, it has the disadvantage that it is not immediately clear to the reader what particular work has been referenced.

It actually does not matter which particular referencing style is used as long as three important considerations are observed:

- the referencing style used throughout the document is consistent;
- all material used or discussed in the text is properly cited;
- nothing is included in the reference list that has not been cited.

This template has a suitable referencing style already set up – you should use it and use the built-in BibTeX system to manage your references. See above for examples of how to cite a reference and look in the `sample.bib` file to see BibTeX references. Remember Google Scholar and other search engines will give you BibTeX references for lots of academic publications. Otherwise, you can easily make up your own based on the examples in that file.

3 Method

seeing LaTeX, or more properly "LaTeX 2ε", is a very useful document processing program. It is very widely used, widely available, stable and free. Famously, TeX, upon which LaTeX is built, was originally developed by the eminent American mathematician Donald Knuth because he was tired of ugly mathematics books**?**. Although it has a learning curve (made much less forbidding by online tools and resources – see below), it allows the writer to concentrate more fully on the content, and takes care of most everything else.

While it can be used as a word processor, it is a *typesetting* system, and Knuth's idea was that it could be used to produce beautiful looking books:

> *LaTeX is a macro package which enables authors to typeset and print their work at the highest typographical quality, using a predefined, professional layout.*[1]

LaTeX has great facilities for setting out equations and a powerful and very widely supported bibliographic system called BibTeX, which takes the pain out of referencing.

Three useful online resources make LaTeX much better:

(1) An excellent online LaTeX environment called "Overleaf" is available at `http://www.overleaf.com` that runs in a modern web browser. It's got this template available – search for a TCD template. Overleaf can work in conjunction with Dropbox, Google Drive and, in beta, GitHub.

(2) Google Scholar, at `http://scholar.google.com`, provides BibTeX entries for most of the academic references it finds.

(3) An indispensable and very fine introduction to using LaTeX called *"The not so short introduction to LATEX 2ε"* by **?** is online at `https://doi.org/10.3929/ethz-a-004398225`.

[1] This is from **?**. Did we mention that you should minimise your use of footnotes?

Browse it before you use LaTeX for the first time and read it carefully when you get down to business.

Other tools worth mentioning include:

- `Draw.io` – an online drawing package that can output PDFs to Google Drive – see `https://www.draw.io`.

4 Results

5 Conclusion

Bibliography

Bedini, I. & Nguyen, B. (2007), 'Automatic ontology generation: State of the art', *PRiSM Laboratory Technical Report. University of Versailles* .

Carpineto, C. & Romano, G. (2012), 'A survey of automatic query expansion in information retrieval', *Acm Computing Surveys (CSUR)* **44**(1), 1–50.

Chiaramella, Y. (2000), Information retrieval and structured documents, *in* 'European Summer School on Information Retrieval', Springer, pp. 286–309.

Galgani, F., Compton, P. & Hoffmann, A. (2012), Combining different summarization techniques for legal text, *in* 'Proceedings of the workshop on innovative hybrid approaches to the processing of textual data', Association for Computational Linguistics, pp. 115–123.

Ge, J., Chen, Z., Peng, J. & Li, T. (2012), An ontology-based method for personalized recommendation, *in* '2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing', IEEE, pp. 522–526.

Goldstein, J., Mittal, V., Carbonell, J. & Kantrowitz, M. (2000), Multi-document summarization by sentence extraction, *in* 'Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization', Association for Computational Linguistics, pp. 40–48.

Gross, B. M. (1964), *The managing of organizations: The administrative struggle*, Vol. 2, [New York]: Free Press of Glencoe.

Roetzel, P. G. (2019), 'Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development', *Business Research* **12**(2), 479–522.

Sarker, A., Mollá, D. & Paris, C. (2013), An approach for query-focused text summarisation for evidence based medicine, *in* 'Conference on Artificial Intelligence in Medicine in Europe', Springer, pp. 295–304.

Soucek, R. & Moser, K. (2010), 'Coping with information overload in email communication: Evaluation of a training intervention', *Computers in Human Behavior* **26**(6), 1458–1466.

A1 Appendix

You may use appendices to include relevant background information, such as calibration certificates, derivations of key equations or presentation of a particular data reduction method. You should not use the appendices to dump large amounts of additional results or data which are not properly discussed. If these results are really relevant, then they should appear in the main body of the report.

A1.1 Appendix numbering

Appendices are numbered sequentially, A1, A2, A3... The sections, figures and tables within appendices are numbered in the same way as in the main text. For example, the first figure in Appendix A1 would be Figure A1.1. Equations continue the numbering from the main text.