

University of Strathclyde
Department of Computer and Information Sciences



Design and Evaluation of an Interactive Topic Detection and Tracking Interface

by
Masnizah Mohd

A thesis presented in fulfilment of the requirements for the degree of
Doctor of Philosophy at the University of Strathclyde

2010

'This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.'

'The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.'

Signed:

Date:

Acknowledgements

My PhD is a challenging journey with wonderful experiences. Having two maternity leaves is one of it and I'm very lucky to be surrounded by truly lovely people. Early in the process of completing this project, it became quite clear to me that a researcher cannot complete a PhD thesis alone. Therefore I would like to thank the following persons for their dedication, prayers and support.

I would like to express my deep and sincere gratitude to my supervisors, Professor Fabio Crestani and Professor Ian Ruthven. I am grateful to them for their commitment, the freedom they gave me to pursue my ideas, the encouragement they provided when I succeeded, the patience they demonstrated when I failed, the wide range of problems they exposed me to, and the direction they consistently provided. Some people wondered as to how I completed my PhD during my supervisor's absence. Although Fabio is currently at University of Lugano, Switzerland, but I never felt like struggling alone. Thank you for being a great supervisor. I would also like to thank Dr. Crawford Revie for his useful comments and encouragement at my annual reviews.

I would like to thank members of the *i-lab* group for making my journey a pleasant one. The administration and support staff in the Department of Computer and Information Sciences (CIS) deserve a huge mention for keeping everything running smoothly. As does my funding bodies, the Ministry of Higher Education, Malaysia and Universiti Kebangsaan, Malaysia (UKM).

Let me also say *thank you* to the following people at CIS: Emma, Christine and Morgan for being wonderful proof readers, research students for your valuable feedback and interest especially during the Researchers' Digest, and the academic staff for your kind help. Thank you to all my friends for the sincere friendship and moral support. My parents and my sisters have always been extremely supportive of me pursuing my educational goals.

Finally to my beloved husband Mohd Zaki, and our children Yusuf Al-Qardhawi, Fatima Az-Zahra and Sarah Ar-Rayyan, all I can say is it would take another thesis to express my deep love for you. Your patience, sacrifice, love and encouragement have upheld my conviction towards pursuing this course, particularly in those days in which I spent more time with my computer than with you. I wouldn't have been able to do this without you.

Abstract

Interactive Topic Detection and Tracking (*i*TDT) is a branch of TDT focussing on aspects of the user interface and user interaction. The importance of user interaction in the real world is the reason why *i*TDT is receiving attention. This particular research project has been motivated by the fact that there has been very little exploration of the usability of the features introduced in the *i*TDT interfaces.

This research investigates the successful components and features of an *i*TDT interface by implementing them into a single interface called Interactive Event Tracking System (*i*Event). The usability of the features introduced in *i*Event, including the usefulness of Named Entity Recognition (NER), were then thoroughly evaluated. The key features of *i*Event - Cluster View, Document View and Term View have enabled journalists to perform various TDT tasks. The user tasks in this study are designed to support the journalists to perform TDT tasks in a way that it is in line with their task. Therefore, this is a ground-breaking study which investigates *journalists* as *i*TDT interface users.

Key findings revealed that *i*Event enables journalists to perform well in TDT tasks. The research also identified potential guidelines for future designs of *i*TDT interfaces for TDT tasks. It was also revealed that, in general, Cluster View is useful and interesting; Document View is effective and interesting; and Term View is helpful. In addition, NER facilitates the Tracking task while keywords are valuable in the Detection task. NER emerges as more valuable in the Tracking task because it can effectively enhance user interactions with the system and brings efficiency to the Detection task. Overall, these findings supported the hypothesis that the use of NER improves *i*TDT as it improves standard TDT.

Contents

Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Interactive Topic Detection and Tracking (<i>i</i> TDT)	2
1.3 Motivation	3
1.4 Research Objective	4
1.5 Overall Layout	5
Chapter 2 Interactive Topic Detection and Tracking	7
2.1 Introduction	7
2.2 Topic Detection and Tracking (TDT)	8
2.3 Topic Detection and Tracking (TDT) Tasks	11
2.3.1 Segmentation	12
2.3.2 Detection	13
2.3.3 Tracking	14
2.4 Interactive TDT Systems	15
2.4.1 Event Organizer	15
2.4.2 TDTLighthouse	16
2.4.3 TimeMine	18
2.4.4 Topic Tracking Visualisation tool	19
2.4.5 Other Systems	20
2.5 User Interface Comparison	23
2.6 Chapter Summary	25
Chapter 3 Methodology	26
3.1 Introduction	26
3.2 Requirement Analysis	27

3.2.1	Review of Interactive TDT Systems	27
3.2.2	Pilot Study	28
3.2.3	Pilot Test	28
3.3	Design and Prototyping	29
3.4	Evaluation	30
3.4.1	Experimental Data	30
3.4.2	Evaluation Participants	32
3.4.3	Experiment Evaluation	32
3.4.4	Evaluation Tasks	34
3.4.5	Evaluation Procedure	39
3.4.6	Training Session	41
3.4.7	Questionnaires	42
3.4.8	System Logging	44
3.5	Chapter Summary	46
Chapter 4 Pilot Study on the use of Named Entity Recognition for <i>i</i>TDT		47
4.1	Introduction	47
4.2	Named Entity Recognition (NER)	48
4.3	Named Entities from Journalism Perspective	50
4.4	Named Entity Recognition (NER) in Interactive TDT (<i>i</i> TDT)	55
4.4.1	Document Representation	56
4.4.2	User Interface	58
4.5	Pilot Study	59
4.6	Methodology	59
4.6.1	Instruction and Online Survey	60
4.6.2	Corpus and Distribution	63
4.7	Results	66
4.7.1	Named Entities Distribution across News Domains	66
4.7.2	The Importance of Named Entities in the News Domains	67
4.7.3	Level of Agreement in the Keywords and News Domains	68
4.8	Discussion	70
4.9	Chapter Summary	73
Chapter 5 <i>i</i>Event User Interface		74
5.1	Introduction	74
5.2	<i>i</i> Event Architecture	75
5.2.1	Document Processing	76
5.2.2	Named Entity Recognition	77
5.2.3	Document Clustering	78
5.3	<i>i</i> Event Interface	79

5.3.1	Cluster View (CV)	82
5.3.2	Document View (DV)	84
5.3.3	Term View (TV)	86
5.4	Example of the use of <i>iEvent</i> in TDT tasks	86
5.4.1	Tracking Task	87
5.4.2	Detection Task	88
5.5	Chapter Summary	89
 Chapter 6 Evaluation of <i>iEvent</i>		91
6.1	Introduction	91
6.2	Organization of <i>iEvent</i> Evaluation	92
6.3	General Findings	92
6.4	Tracking Task	97
6.4.1	Overall Opinions	97
6.4.2	Reporting Task	103
6.4.3	Features	106
6.4.3.1	Useful	106
6.4.3.2	Effective	109
6.4.3.3	Helpful	112
6.4.3.4	Interesting	116
6.5	Detection Task	119
6.6	Discussion	124
6.7	Chapter Summary	127
 Chapter 7 Evaluation of the use of Named Entity Recognition in <i>iEvent</i>		128
7.1	Introduction	128
7.2	Evaluation Methodology	129
7.3	General Findings	130
7.4	Tracking Task	133
7.4.1	Overall Opinions	133
7.4.2	Reporting Task	136
7.4.3	Profiling Task	136
7.4.4	Features	139
7.5	Detection Task	145
7.6	Discussion	148
7.7	Guidelines for the designs of the <i>iTDT</i> interfaces	150
7.8	Chapter Summary	153

Chapter 8 Conclusion and Future Work	155
8.1 Introduction	155
8.2 Contributions	156
8.2.1 Novel <i>i</i> TDT Interface	157
8.2.2 Proper Evaluation on <i>i</i> TDT	158
8.2.3 The Importance of Named Entity Recognition (NER) in <i>i</i> TDT	158
8.3 Future Works	158
8.3.1 Novel <i>i</i> TDT Interface	159
8.3.2 Proper Evaluation on <i>i</i> TDT	160
8.3.3 The Importance of Named Entity Recognition (NER) in <i>i</i> TDT	161
8.4 Chapter Summary	161
Bibliography	163
Published Work	170
A Resources relating to the work described in Chapter 3	171

List of Figures

2.1	TDT Topic	9
2.2	Event Organizer user interface	16
2.3	TDTlighthouse user interface	17
2.4	TimeMine user interface	18
2.5	Topic Tracking Visualisation tool	19
2.6	Press Display interface	21
2.7	Paid Content interface	22
2.8	Google Fast Flip interface	23
3.1	The incremental design methodology	27
3.2	Simulated situation	35
3.3	Topic Detection task	36
3.4	Experimental design	38
3.5	Likert scale taken from the <i>Tracking</i> questionnaire	43
3.6	Semantic differentials	44
3.7	Log system	45
4.1	Hypothetical structure for a News Schema (Dijk, 1983)	53
4.2	News report for topic <i>Pope Visit Cuba</i> from CNN	55
4.3	Online survey main interface	60
4.4	Online survey	62
4.5	User interface approach	72
5.1	iEvent architecture	75
5.2	Rules in ANNIE for Company	78

5.3	Single-pass algorithm	78
5.4	Keywords Setup (Setup 1)	80
5.5	Named entities Setup (Setup 2)	81
5.6	Size and density of the clusters	82
5.7	Cluster labelling of the two approaches	83
5.8	Top ten terms of the two approaches	84
5.9	Document histogram for topic <i>Cable Car Crash</i> (Topic 20019)	84
5.10	Document content of the two approaches	85
5.11	Histogram and the timeline for named entity <i>Italy</i> (Topic 20019)	86
5.12	Histogram and the timeline for named entity <i>James Earl Ray</i> (Topic 20056)	86
6.1	Percentage of the news networks tools used	96
6.2	Percentage of successful Tracking task	101
6.3	Boxplot of <i>how much was written</i> (scale 1-5) by each type of participant in the Tracking task	104
6.4	Percentage of correct news written	105
6.5	Boxplot on the <i>usefulness</i> (Scale 1-5) of the features for type of participants	108
6.6	Boxplot on the <i>effectiveness</i> (Scale 1-5) of the features for type of participants	111
6.7	Boxplot on the <i>helpfulness</i> (Scale 1-5) of the ‘TV: histogram with the time line’ feature for Topics 1-7	114
6.8	Boxplot on the <i>helpfulness</i> (Scale 1-5) of the features for type of participants	115
6.9	Boxplot on the <i>interestingness</i> (Scale 1-5) of the feature for type of participants	118
6.10	Percentage of successful Detection task	120
7.1	Boxplot of participants’ topic interest (scale 1-5) after using <i>iEvent</i> across setups in the Tracking task	131
7.2	The percentage of the participants’ preferred setup for the given tasks	133
7.3	Boxplot of participants’ opinion (scale 1-5) for the setups in the Tracking task	134
7.4	Percentage for the type of keywords	137
7.5	The percentage for the type of keywords across setups	138
7.6	Boxplot on the <i>usefulness</i> (scale 1-5) of the ‘CV: cluster labelling’ feature across setups	139
7.7	Boxplot on the <i>effectiveness</i> (scale 1-5) of the ‘CV: cluster labelling’ feature across setups	141

7.8	Boxplot on the <i>helpfulness</i> (scale 1-5) of the ‘CV: top terms’ feature and ‘TV: keyword approach’ feature across setups	142
7.9	Boxplot on the <i>interestingness</i> (scale 1-5) of the ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups	144
7.10	Boxplot on <i>Easy to Detect</i> (Scale 1-5) of the setup for the Detection task ...	146

List of Tables

2.1	Description and an example of named entities (NE)	11
2.2	Comparison of <i>i</i> TDT features	24
3.1	TDT2 and TDT3 topics	31
3.2	TDT standard performance measures	33
3.3	Methods for calculating the effectiveness measures in TDT	33
3.4	Topics in the Tracking task (F1-measure)	34
3.5	Clusters in the Detection task (F1-measure)	34
3.6	Topics in the Tracking task	38
3.7	Clusters in the Detection task	39
3.8	Interaction tag	45
4.1	Documents Distribution	63
4.2	Documents across news domains	65
4.3	Named entities distribution across news domains	66
4.4	Importance of named entities across news domains	68
4.5	Overlap values for documents across news domains	70
5.1	Type of named entities	81
6.1	The mean for topic familiarity and topic interest	94
6.2	Topic familiarity and topic interest	94
6.3	Topic interest (after) across setups	95

6.4	Percentage of the participant opinions of the news network tools used	96
6.5	Percentage of the participant opinions of <i>iEvent</i>	98
6.6	Percentage of the participant opinion (easy) across setups	98
6.7	Percentage of the participant opinion (relaxing) across setups	99
6.8	Percentage of the participant opinion (interesting) across setups	102
6.9	Comparison of news written on topic Mobil-Exxon Merger between the student and the journalist	104
6.10	Percentage of participants who perceived the features of <i>iEvent</i> as <i>useful</i> in the Tracking task	107
6.11	‘CV: cluster labelling’ feature across setups perceived as useful	109
6.12	Percentage of participants who perceived the features of <i>iEvent</i> as <i>effective</i> in the Tracking task	110
6.13	‘CV: cluster labelling’ feature across setups perceived as effective	112
6.14	Percentage of participants who perceived the features of <i>iEvent</i> as <i>helpful</i> in the Tracking task	113
6.15	‘CV: top terms’ and ‘TV: keyword approach’ features across setups perceived as helpful	116
6.16	Percentage of participants who perceived the features of <i>iEvent</i> as <i>interesting</i> in the Tracking task	117
6.17	‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups perceived as interesting	119
6.18	Percentage of participants’ opinion on <i>easy to detect</i>	121
6.19	Percentage of the participant opinions on the ease to detect a topic across setups	121
6.20	The percentage of <i>iEvent</i> features used in Detection task	122
6.21	The significant value for the frequency of the features across setups (Mann-Whitney Test)	123
6.22	The ratio of each feature across participants’ opinion in Tracking task	124
6.23	The comparison of each feature in facilitating the TDT tasks	126
7.1	The differences of the two setups	130
7.2	Topic interest (after) across setups	132
7.3	Percentage of the participant opinions of Setup 2	134
7.4	The frequency for type of keywords across topics	137
7.5	‘CV: cluster labelling’ feature across setups perceived as useful	140
7.6	‘CV: cluster labelling’ feature across setups perceived as effective	141
7.7	‘CV: top terms’ and ‘TV: keyword approach’ features across setups	143
7.8	‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups	145
7.9	Percentage of the participant opinions on the ease to detect a topic across setups	147

7.10	The significant value for the frequency of the features with NER across setups (Mann-Whitney Test)	147
7.11	The comparison of features with approaches in facilitating the TDT tasks ..	149
7.12	Comparison on the mean of click for each successful task between setups ..	150
7.13	Guideline for the designs of the <i>i</i> TDT interfaces	151
7.14	Comparison of the approaches across TDT task	153

Chapter 1

Introduction

1.1 Introduction

The explosive growth and dynamic environment of digital information threatens to overwhelm human attention, thus raising new challenges for information retrieval (IR) technology. How can we track and detect dynamic information such as news using IR techniques? Teevan (2007) in her study on how people re-find information on the Web uses the term *dynamic information* to refer to *any information that has changed in any way*. There are many IR systems publicly available that aim to help users become aware of the most current news on the Web. For example, services such as GoogleNews¹, NewsInEssence² at the University of Michigan offer a tracking service in which users receive an email when new articles about their subject of interest become available. Such services track information updates at the document level. Information professionals such as journalists often rely on tools such as Rich Site Summary (RSS) news feeds to keep track of the most current information and events. Thus there is an increasing need

¹ <http://news.google.com>

² <http://www.newsinesence.com>

for automatic techniques to analyse, present and visualise news to users in a meaningful and efficient manner.

Dynamic information is the main topic dealt with in the area of research known as Topic Detection and Tracking (TDT). Research in TDT aims to effectively retrieve and organise broadcast news (speech) and newswire stories (text) into groups of events. The majority of TDT research and evaluation has been on the system performance without any user involvement. TDT is part of Text Retrieval Conference (TREC). In recent years, much work has been done in TREC and the TDT domain to investigate methods for automatically organising news stories. This research is part of TREC which involves user and task based evaluation and will be discussed in Chapter 3. Very few TDT researchers have started working on user interfaces and user interaction. The priority of their work has been restricted to system performance without proper evaluation from user perspective.

1.2 Interactive Topic Detection and Tracking (*i*TDT)

The initial motivation for research in TDT was to provide a core technology for a system that would both monitor broadcast news and alert an analyst to new and interesting events occurring in the world. Analysts are keen to track and in particular to know the latest news about a story from a huge volume of information that arrives daily. There is an attempt in TDT research to focus on user interaction, user evaluation and user interfaces as a way to visualise and represent news in a meaningful way. The term interactive TDT (*i*TDT) is used for the first time in this thesis, to refer to the TDT works which focus on these aspects. It is important to provide a means for people such as journalists to understand and interpret what is happening in the news. There is still TDT research active where the researchers in this area have focused on developing algorithms for better TDT performance and the evaluation of these algorithms is the main activity in TREC evaluation. Few TDT researchers have investigated techniques such as information visualisation and automatic timelines to support users with a dynamic and

interactive use. Very few researchers have worked on interfaces and user interaction for TDT. According to Shneiderman (1997), an effective interface should be well designed and generate a positive feeling of success, competence, mastery, pleasure and clarity in the user community. TDT research is still continuing: one of the focuses now is on *i*TDT and this work is in this direction.

The main focus of this thesis is to design an interactive TDT (*i*TDT) interface and to evaluate the effect of the interface on the effectiveness of user performance. The evaluation also includes the usage of Named Entity Recognition (NER) on the interface. NER received considerable attention in TDT because detection of named entities enables a system to characterise and detect events in documents. Named entities include information units like names, including person, organisation, location names, and numeric expressions including time, date, money and percent expressions. The *Person* and *Location* type of named entities will provide the reader with information on *Who* is involved and *Where* the event is. It helps them to understand better what the event is about via the use of NER and helps journalists to perform their TDT tasks.

1.3 Motivation

Most previous research in TDT has concentrated primarily on the design and evaluation of algorithms to carry out TDT tasks such as stream segmentation; link detection; story detection; story tracking, in a batch way, fully automatically and without the need for user interaction. Evaluation, carried out for a number of years in the context of TREC, has always been done exclusively laboratory-style, without any user involvement. The starting point of the work reported in this thesis is that TDT is very much an interactive task, since the combination of subtasks that make up the TDT task are very difficult to study in isolation, given their interdependence. Studying TDT from a user interaction perspective enables us to view the TDT task in its entirety. For example it is very difficult to segment a news stream and to define the news story, but a user can carry this

out in a very effective way. Thus, there is an increasing need in TDT for user interfaces to help the users to analyse and inspect news effectively.

Recently, increasing effort has been devoted to designing user interfaces to improve TDT systems by investigating not just the interaction aspect but also a user and task oriented evaluation. I believe that interfaces play a vital role in *i*TDT and this has motivated me to design a new interface for interactive TDT that is meant to support the user in all the tasks related to TDT. The importance of user interaction in the real world is the reason why *i*TDT is receiving more attention.

It was clear that a well designed *i*TDT interface is important to guide users in performing the TDT tasks. Designing such an interface should incorporate the best and most successful components or features. In addition, one important element of TDT research is the realisation of the importance of Named Entity Recognition (NER) in all tasks related to TDT (Allan et al., 2005). TDT focuses on event-based news organisation which requires information at the very least on *what* happened, *where* it happened, *when* it happened, and *who* was involved. The use of NER in TDT seems to be an advantage. Therefore an *i*TDT interface should be able to display named entities and use NER. Although proved in TREC experimentation, it is not clear if the use of NER really improves the effectiveness of TDT. This is because NER has been used but only from the algorithmic perspective.

The user and task oriented evaluation will then enable us not only to test the effectiveness of the novel interface for *i*TDT but also the interactive use of NER in facilitating the journalists to perform TDT tasks.

1.4 Research Objective

This thesis addresses issues of the design of an *i*TDT interface that uses features that are effective in allowing the professional such a journalists to perform the TDT tasks. It also

highlights the interaction between journalists and a TDT system using NER applied on user interfaces in performing the TDT tasks.

I investigate named entity usage from the journalism perspective and present it in an interactive way on the user interface. The research goals are to present a well designed *i*TDT interface and to evaluate the usability of the features introduced. The three main aims of the thesis are as follows:

- a. To present the main guidelines for the design of *i*TDT interface:
What are the elements of the design of an interface that aims to facilitate journalists to perform TDT tasks?
- b. To evaluate the components and features of *i*TDT interface:
It is important to identify the effective and the useful components of an *i*TDT interface and what TDT tasks these components are facilitating.
- c. To evaluate the use of Named Entity Recognition (NER) in *i*TDT:
Is the use of NER improving *i*TDT in the same way as it has been shown to improve standard TDT? What are the TDT tasks facilitated through the use of NER?

1.5 Overall Layout

This thesis is divided into eight chapters:

Chapter 1: Introduction

This chapter provides the background and motivates the work described in this thesis.

Chapter 2: Interactive Topic Detection and Tracking

Chapter 2 presents an overview of TDT, *i*TDT and interfaces. It discusses the definition of Topic and Event in TDT and three technical tasks in TDT such as Segmentation, Detection and Tracking. This chapter aims to identify the components and features for a *i*TDT interface.

Chapter 3: Methodology

This chapter discusses the methodology of this work and describes each methodology component.

Chapter 4: Pilot Study on the use of Named Entity Recognition for *i*TDT

This chapter discusses NER in general and focuses on its use in TDT and *i*TDT. This chapter also discusses the findings of a pilot study which aims to understand named entities from a journalism perspective and to investigate how useful it is in *i*TDT.

Chapter 5: *i*Event User Interface

This chapter presents the design and implementation of a prototype system called *i*Event (Interactive Event Tracking System). It describes the architecture and the components of the *i*Event user interface.

Chapter 6: Evaluation of *i*Event

This chapter discusses the evaluation of the usability of *i*Event. It presents the results of the user experiment and the discussion of findings which consist of the general findings, the participants' performance in the TDT task and the comparison of their performance in both tasks. This chapter aims to identify the effectiveness of the components and features in *i*Event and to associate it with the TDT tasks.

Chapter 7: Evaluation of the use of Named Entity Recognition in *i*Event

In this chapter, I discuss the evaluation of NER applied on the interface. This chapter aims to identify the features of *i*Event and TDT tasks which will perform better with the use of NER.

Chapter 8: Conclusion and Future Work

This chapter discusses several remaining issues, novel contributions and indicates some future research directions.

Chapter 2

Interactive Topic Detection and Tracking

2.1 Introduction

This chapter provides the background for the research described in this thesis and creates a context within which the work is situated. It contains an introduction to Topic Detection and Tracking (TDT) with explanation of the TDT tasks. The focus of this chapter is the review of the related works on interactive TDT (*i*TDT), with a view to identifying the good components and features for a well designed *i*TDT interface.

With an increasing amount of information coming from different sources such as newspapers, radio, television and more recently the Web, a proper organisation of news is vital. A research area such as TDT, monitors news streams and organises broadcast news (speech) and newswire stories (text) into groups of events. This news stream may or may not be pre-segmented into stories, and the events may or may not be known to the system where the system may or may not be trained to recognise specific events.

This leads to the definition of three technical tasks to be addressed in the TDT study. These are: the tracking of known events, the detection of new events, and the segmentation of a news source into stories. Thus, it would be desirable for an intelligent system to automatically detect significant events from large volumes of news stories; present the main content of events to the user in a summarised form with multiple levels of concept; alert the onset of novel events as they happen and track events of interest based on user-given sample stories. This is the goal of Topic Detection and Tracking (TDT).

2.2 Topic Detection and Tracking (TDT)

TDT is a body of research and an evaluation paradigm that focuses on event-based news organisation (Allan, 2002). TDT was originally funded and supported by the Defence Advanced Research Projects Agency (DARPA), but it is now under the control of the Translingual Information Detection, Extraction and Summarisation (TIDES) program. In the majority of years the evaluation attracts roughly 11 participants including its founding members; University of Massachusetts, Carnegie Mellon University and Dragon Systems; plus other important participants such as IBM Watson and the University of Maryland. After 7 open and competitive annual evaluations from 1998 to 2004, TDT has become quite mature. However, after having completed one pilot study (TDT 1997) and continued through TDT 1998 to TDT 2004, it is still an active area of research.

In the initial TDT study, the notion of a topic was limited to be that of an event, which means something that happened at some specific time and place. Then the definition of a topic was broadened to include other events and activities that are directly related to it. The TDT definitions of a topic and an event that are still used in current evaluations were agreed upon during the second TDT evaluation in 1998, and are defined as follows (Allan, Papka & Lavrenko, 1998):

- a. A topic is a seminal event or activity along with all directly related events and activities.
- b. An event is something that happens in a specific time and place (specific elections, accidents, crimes and natural disasters are examples of events).
- c. An activity is a connected set of actions that have a common focus or purpose (specific campaigns, investigations, and disaster relief efforts are examples of activities).

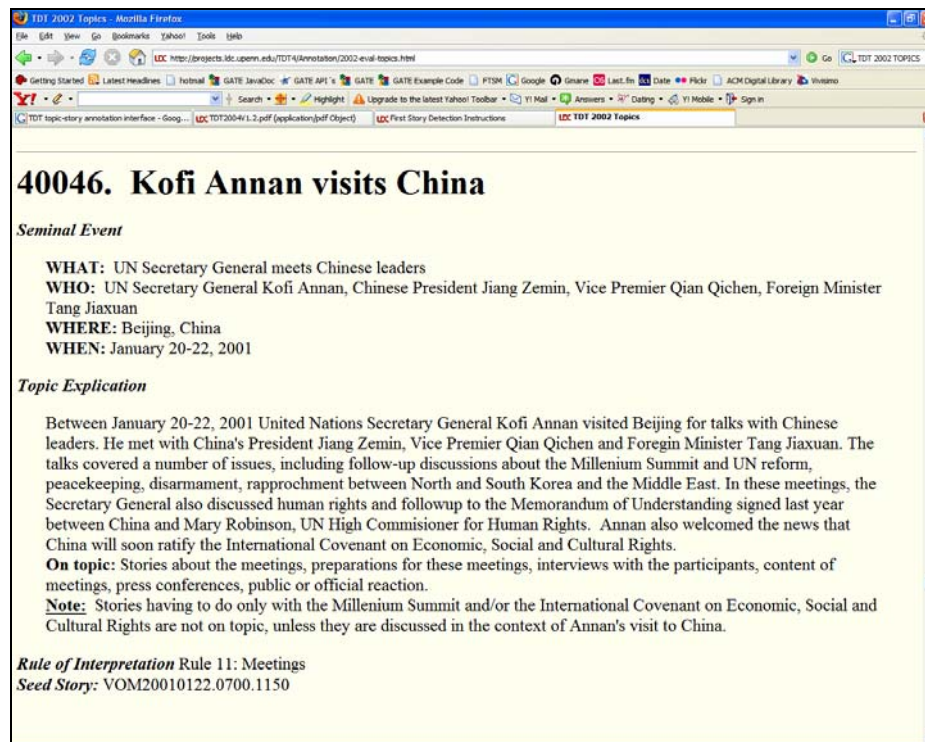


Figure 2.1: TDT Topic

Figure 2.1 shows an example of a TDT topic on *Kofi Annan visits China* with the seminal event highlighting on the *Who*, *What*, *Where* and *When*.

Topics in TDT are also specific where it differentiates between different instances of *typhoon* in the general topic of *typhoon*. For example:

- i. 4th of November 1995, *Super Typhoon Angela* Hits Philippine Heartland
- ii. 24th of July 2003, *Typhoon Imbudo* slams into China
- iii. 27th of September 2005, *Typhoon Damrey* makes landfall in Vietnam
- iv. 19th of November 2007, *Cyclone Sidr* second strongest to hit Bangladesh
- v. 26th of November 2007, *Typhoon Mitag* hits Philippines

A TDT system draws a distinction between events and topics. As Yang et al. (1999) note, "*The USAir- 427 crashes is an event but not a topic, and airplane accidents is a topic but not an event*". Studies by various authors (Makkonen et al., 2004; Yang et al., 1999; Cieri et al., 2002, Yang et al., 2002) have emphasised the importance of having a better understanding and a clear definition of an event. Indeed, it appears that event is a term which is easy to understand at the intuitive level but hard to define precisely. An event comprises at the very least *what* happened, *where* it happened, *when* it happened, and *who* was involved. In the Information Extraction (IE) field, Message Understanding Conferences (MUC) have presented an approach to define an event using a Scenario Template (ST) task. The main goal of ST is to extract pre-specified event information and relate the event information to a particular organisation, person, or other entities involved in the event (Allan, 2002).

The identification and extraction of named entities relates to information about Date (*when*), Location (*where*), Person (*who*) and Organisation (*what/who*). It is crucial in TDT because detection of named entities allows the process of identifying an event in a document to occur. Thus it is important to have a clear understanding of named entities in TDT. As a result, I conclude that an event is defined as a unique circumstance or a condition that involves the integration of the W's elements as shown in Table 2.1.

NE	Description	Example
Who	This element refers to the actor or person that takes part in the event.	Alex Salmond won the Gordon constituency in the 2007 Scottish Parliament election.
Where	This element refers to the places or location that the event takes place.	Scotland was the location for the 2007 Scottish Parliamentary election.
When	This element refers to the date or time that the event takes place.	3rd May 2007 was the election day.
What	This element refers to the subject, occasion, body or activity that involved in the event.	The Scottish National Party (SNP) won the 2007 Scottish Parliament election.

Table 2.1: Description and an example of named entities (NE)

However, simple uses of Named Entity Recognition (NER) seem to be not very helpful for improving the performance of a TDT system. Therefore I try to address issues on how to effectively make use of NER in TDT by focusing on works related to document representation and user interface design. In the next section, I explain the three technical tasks to be addressed in TDT.

2.3 Topic Detection and Tracking (TDT) Tasks

Topic Detection and Tracking (TDT) is a research program investigating methods for automatically organising news stories by the events that they discuss. TDT includes several evaluation tasks, each of which explores one aspect of the organisation of a continuous stream of news, including: a) splitting the stream into stories that are about a single topic (stream segmentation); b) gathering stories into groups that each discuss a single topic (link detection); c) identifying the onset of a new topic in the news (first story detection); and d) exploiting user feedback to monitor a stream of news for

additional stories on a specified topic (story tracking). The goal of a TDT system is to monitor a stream of broadcast news stories, and to find the relationships between these stories based on the events that have been described.

2.3.1 Segmentation

Segmentation is the task of breaking a broadcast news stream into its constituent news stories. The necessity of this task is related to the added difficulty of working with broadcast radio and television transmissions. Unlike written sources of news, which contain title, paragraph and story boundary information, a broadcast news transcript or closed caption material will not contain any mark-up indicating where stories begin and end in the data stream. This has prompted an entirely new opportunity in research that requires systems to automate the story segmentation process. Thus, TDT systems must incorporate more robust filtering technologies that can tackle noisy input due to segmentation errors (Allan et al., 2002). For example, a missed story and additional errors contained in Automatic Speech Recognition (ASR) system output, such as lack of capitalisation, and errors due to pronunciation similarity between different word forms such as *icing* and *I sing*. Much of the Linguistic Data Consortium (LDC)'s work in creating the TDT corpora was concerned with adding boundary information to automatic ASR output and closed-caption transcripts. Segments in TDT text must also be classified as one of the following: a news story, a miscellaneous news item such as reporter chat or advertisements; and untranscribed text containing incomplete stories where there is not enough information present in the text to identify its topic (Cieri et al., 2002). These human-identified topic boundaries are then used to evaluate the performance of TDT segmentation systems. The TDT community has also investigated the impact of automatic segmentation errors on other TDT tasks, where it has found that segmentation has little effect on tracking tasks, but does dramatically affect the impact of various detection tasks (Allan, 2002).

2.3.2 Detection

Detection is the task of identifying similar (on-topic) and dissimilar (off-topic) news stories in the news stream. The detection task is characterised by the lack of knowledge of the event to be detected. Detection can be further subdivided into new event detection, cluster detection, and link detection tasks.

a. (Online) New Event Detection (NED)

(Online) New Event Detection (NED) is the task of recognising seminal events as they arrive on the data stream. In TDT 1999 – 2002 this task was referred to as First Story Detection, however, the TDT community have reverted back to calling it by its original task name as it appeared in the 1997 pilot study. In all evaluations, the task definition remains the same; to find the document that is the first to discuss a breaking news story for each event in the collection. This is an online filtering task so the system can only make this decision (first story or not a first story) for the current document by considering only those documents that it has seen so far on the input stream.

b. Cluster Detection

Cluster Detection has been referred to as either Event Detection or Retrospective Event Detection in previous TDT evaluations. The task definition for an event detection system is to divide the data stream into clusters of related events by considering all the documents in the TDT collection rather than just those that occur before the current document in the input stream, as in the case of online new event detection. This task has proved to be considerably more popular than new event detection due to the similarity of this technology with previous research efforts such as clustering-based Text Retrieval Conference (TREC) tasks.

c. Story Link Detection

Story Link Detection is the task of classifying a pair of news stories as on-topic (belong to the same topic) or off-topic (belong to different topics). The TDT initiative has

emphasised the importance of this task as it is *a core technology for all other tasks* (Allan, 2002). This claim is easily understood since all Information Retrieval (IR) and filtering systems are concerned with the determination of document similarity. It is hoped that the advancement of other TDT tasks may be possible by refining story link detection in TDT.

2.3.3 Tracking

Tracking is the task of finding all subsequent stories in the news stream that relate to a certain known event represented by the first n sample stories on that event. The tracking task associates incoming stories with events known to the system. An event is defined by its association with stories that discuss a particular event. Thus each target event is characterised by a list of stories which discuss it. Each successive story must be classified according to whether or not it discusses the target event in the tracking task. Therefore, the study corpus is divided into two parts, with the first part being the training set and the second part being the test set. Each of the stories in the training set is flagged in order to highlight whether it discusses the target event or not, and these flags (and the associated text of the stories) are the only information used for training the system to classify the target event correctly.

It is similar to the TREC information filtering task. Each begins with a representation of a topic and then monitors a stream of documents, making decisions as they arrive. Each document is assigned a score for that topic and, if the score is high enough, it is retrieved. Filtering simulates interacting with the user to supervise the process, whereas tracking operates as if the user were not there. Systems may be adaptive in that they *guess* that a story is on topic, but they do not receive human confirmation that they were correct.

2.4 Interactive TDT Systems

TDT researchers have attempted to build better document models, developing similarity metrics or better document representations. This led to a series of research efforts that concentrated on improving document representation by applying Named Entity Recognition (NER) (Yang et al., 1999; Makkonen et al., 2004; Kumaran & Allan, 2004; Kuo et al., 2007). Then a few researchers started to move from the laboratory style of experiment to the interactive TDT mainly focusing on graphical user interface (GUI). Event Organizer (Allan et al., 2005), TDTLighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000) and Topic Tracking Visualisation tool (Jones & Gabb, 2002) are an example of TDT works that investigate certain approaches to improving TDT system performance using GUI. I review these works by discussing the features and the approaches used and how it motivates this work.

2.4.1 Event Organizer

Event Organizer (Allan et.al, 2005) is a TDT system that aims to organise a constantly updating stream of news articles by the events that are discussed in the stories. It does not only focus on the cluster detection technology but also on the user interface employing a Document View with the timeline. This is one of the best features in the user interface since date is an important indication of when the event occurs. Through the interface, users are allowed to correct the system's errors by removing stories from clusters and creating new clusters in their profile, as shown in Figure 2.2.

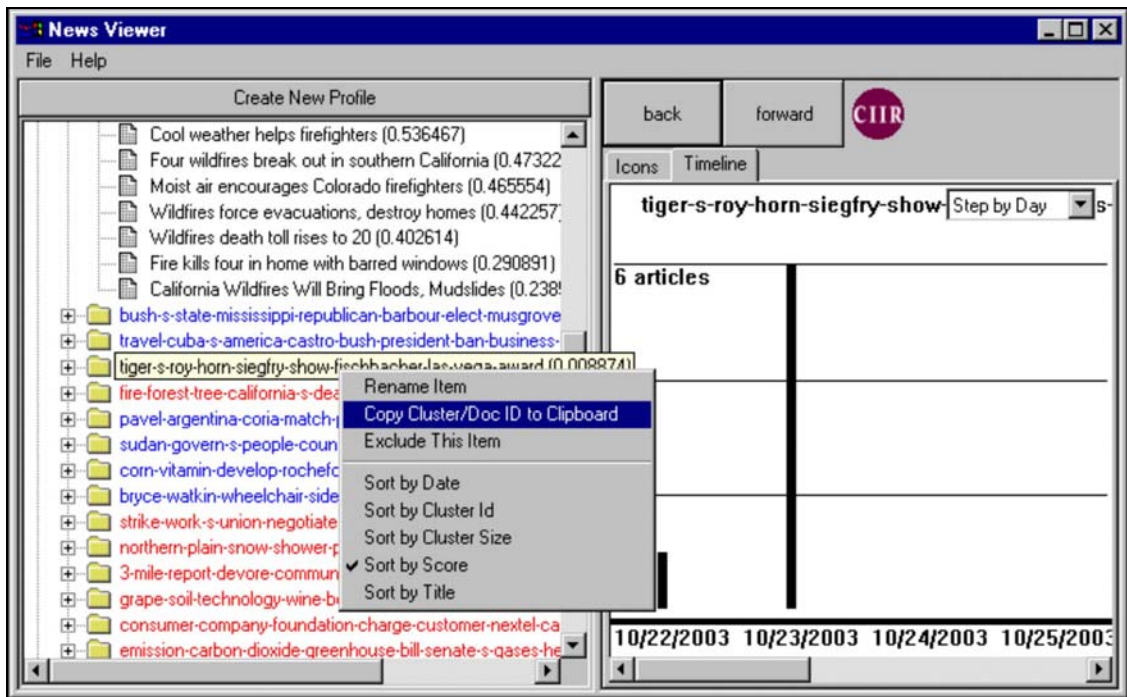


Figure 2.2: Event Organizer user interface

The user can create a profile that captures clusters of interest. A profile is a representation of a folder that contains all clusters that matched the query. Clusters are named with the ten most highly weighted terms or named entities occurring in the stories inside the cluster.

2.4.2 TDTLighthouse

Another related work is the TDTLighthouse (Leuski & Allan, 2000) that has been designed for presenting results of a search session to the user, as shown in Figure 2.3. It provides not only a typical ranked list search result, but a visualisation of inter-document similarities in two or three dimensions. The visualisations present the documents as spheres floating and position them in proportion to their inter-document similarity. If two documents are very similar to each other, the corresponding spheres will be closely located and the spheres that are positioned far apart indicate very different page content.

Visualising the cluster in sphere form is the strength of this work since this feature helps the user to understand news in a relatively fast and efficient manner, thus enabling them to focus on the relevant documents more accurately.

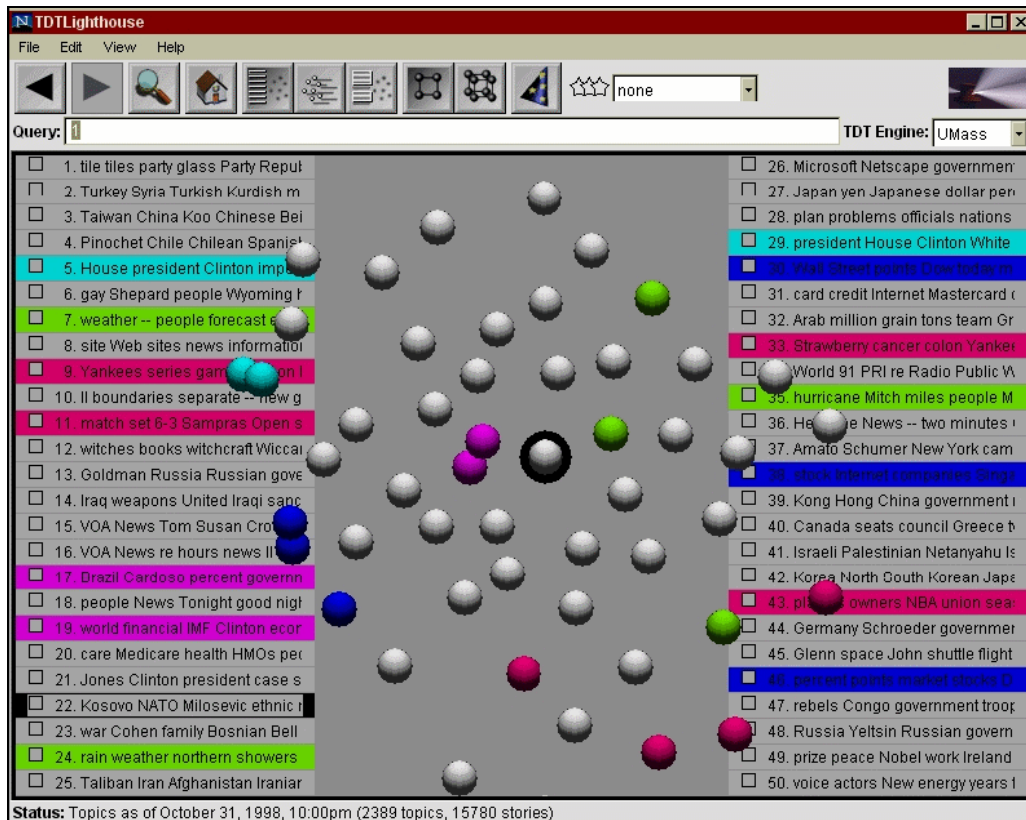


Figure 2.3: TDTLighthouse user interface

This motivates me to visualise the relevance of the documents contained in a cluster based on the size and density. For example, a cluster of large size and high density contains a high number of documents and that the documents are in a short period of time. This might indicate the cluster containing an important event. In TDTLighthouse, users have to judge the relevance of documents in the context their information needs by looking through the titles of document. I believe it will be better if there is an approach to present and label the relevant documents using named entities.

2.4.3 TimeMine

A prototype system called TimeMine (Swan & Allan, 2000) is a TDT system that aims to detect, rank and group semantic features based on their statistical properties. It automatically generates an interactive timeline displaying the major events and uses it as a browsing interface to a document collection, as shown in Figure 2.4. It restricts the amount of information on the interface by presenting the most significant and important information to the user. The timelines are the best features in this work since it provides an effective form of presentation and a very fast graphical overview of the information that a corpus contains. I believe timelines are useful and have been motivated to provide this feature in my work.

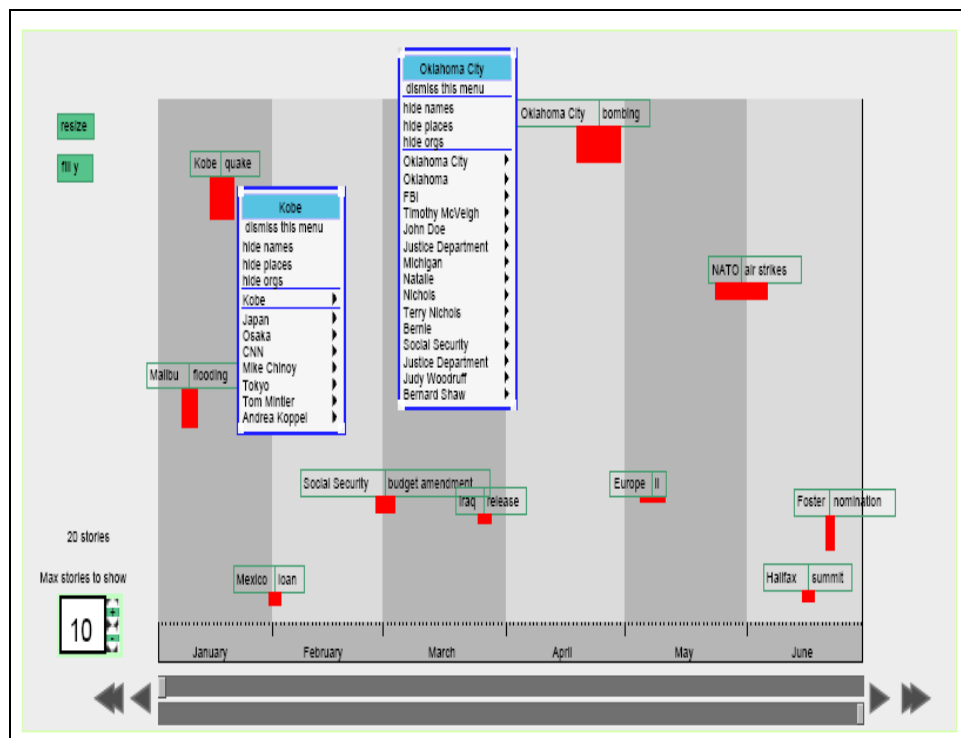


Figure 2.4: TimeMine user interface

2.4.4 Topic Tracking Visualisation tool

In (Jones & Gabb, 2002), an interactive graphical visualisation tool is presented for use in TDT algorithm development. The system uses colours to show the results of the TDT system in relation to some *ground truth*. For example, on-topic stories are shown in green, misses are shown in red, and false alarms are shown in blue, as shown in Figure 2.5. Thus, this work is clearly directed towards the design of an interface for measuring TDT performance through the use of visualisation. In fact, this interface allows the user, for example, to easily identify the changes in the false alarm rate if the threshold changes. The system enables easy selection of system parameter settings with interactive graphical display of the results (which can be pre-computed), as well as the standard tracking measures mentioned. For example, the user could change the setting of the threshold, the amount of training stories known, or view the topic, the words and the statistics.

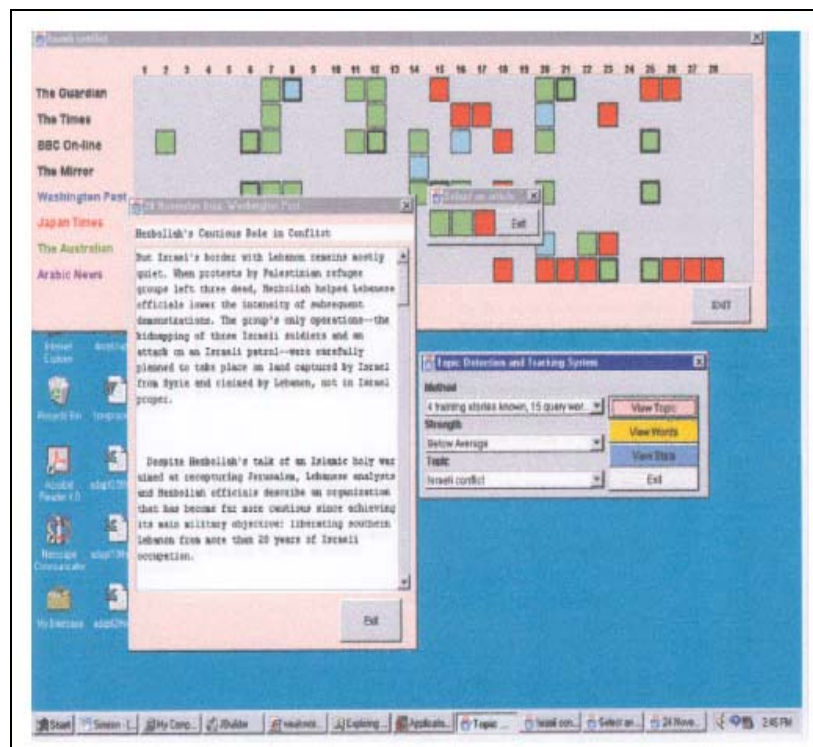


Figure 2.5: Topic Tracking Visualisation tool

Visualisation is a very powerful way of representing large amounts of information for rapid interpretation (Leuski & Allan, 1998). Timelines are a well known interface, simple and intuitive for most people to use. There has been very little exploration of how cluster visualisation and timelines could be effectively used together. It was obvious that very few of these related works have combined the Document View, Cluster View and Term View for interactive TDT. None of the works discussed above evaluated the usability of the features introduced in the interfaces and almost none of them evaluated the effect of the interface on the effectiveness of the user performance.

2.4.5 Other Systems

There are other systems that people use for doing *i*TDT. These systems cannot really be defined as *i*TDT because they don't really carry out any of the TDT tasks but still people use them for news monitoring and searching in systems such as PressDisplay.com³, PaidContent.org⁴ and Google Fast Flip⁵. Therefore I do not include these systems in the requirement analysis towards the design and implementation of the interface as reported in Chapter 3 (Section 3.2). These systems were introduced as a result of the post - evaluation interview on the participants' previously used news network tools which is discussed in Section 6.3 (General Findings). They refer to these systems as the systems that they use for the task related to TDT.

PressDisplay is a web-based portal which provides online access to over 1,000 newspapers and magazines from more than 76 countries in more than 38 languages, as shown in Figure 2.6. It was released in 2003 and the publications in it are displayed in their original format and can serve as a formal reference. It has a structured and simple user interface, easy to use navigation tools and contains a timeline feature where readers can gain access to newspapers for a week. It provides readers with comprehensive

³ <http://www.pressdisplay.com/pressdisplay/viewer.aspx>

⁴ <http://paidcontent.org/>

⁵ <http://fastflip.googlelabs.com/>

SmartNavigation software that processes newspaper files as they arrive from the publishers and intelligently recognises images and articles. These files are then presented to readers in an interactive format, allowing readers even greater functionality from their selected publication. SmartNavigation also provides readers with advanced digital tools such as Article View, Table of Contents, Article Linking, URL/Email linking, Language Translation, Sound Integration and RSS Feeds. PressDisplay also allows personalization, with features such as *My Monitor*, *My Preferences* and *My newspaper* based on their topic of interest.



Figure 2.6: Press Display interface

PaidContent.org is an online media hub that covers news, information and analysis of the business of digital media as shown in Figure 2.7. It was launched by a journalist,

Rafat Ali in April 2002. The interface is not as structured and interactive as PressDisplay but it has clear categories of content. It also provides the reader with a job service where users can search jobs based on *What* and *Where*. Readers can view the latest and the popular news with features such as RSS Feeds.



Figure 2.7: Paid Content interface

Finally **Google Fast Flip** is a web application that combines the qualities of print and the Web, with the ability to *flip* through pages online as quickly as flipping through a magazine, as shown in Figure 2.8. The stories are grouped by categories, such as Entertainment, Business, Opinion, Politics and Most Viewed. Readers can flip through stories quickly by simply pressing the left- and right-arrow keys until they find one that catches their interest. Clicking on the story takes them directly onto the publisher's

website. It also allows personalization, enabling readers to follow friends and topics, discover new content and create their own custom magazines around searches.

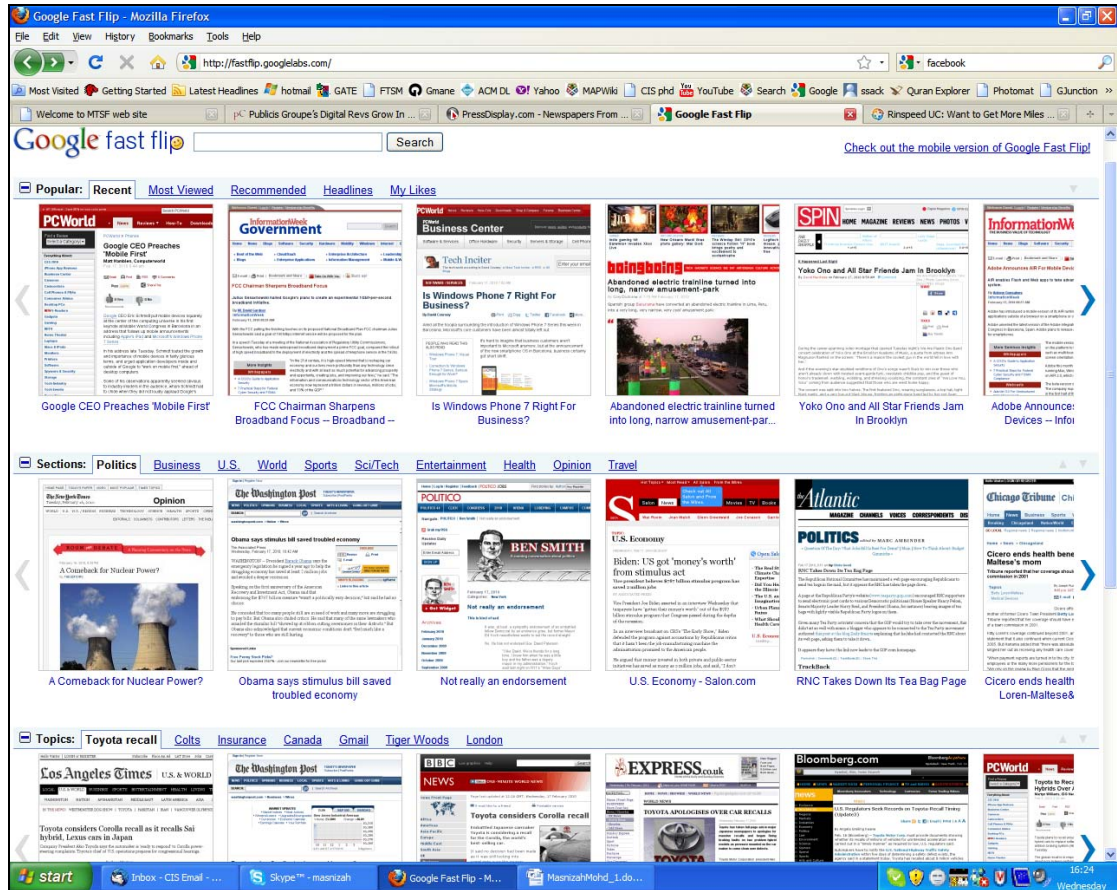


Figure 2.8: Google Fast Flip interface

2.5 User Interface Comparison

The reviewed works on interactive TDT enabled me to identify the similarities and differences of the components and features used, as shown in Table 2.2. This comparison does not include *Other Systems* (section 2.4.5) because they are not really TDT systems.

	Document View (DV)	Cluster View (CV)	Term View (TV)
Event Organizer	<ul style="list-style-type: none"> • Story profile (important terms and document list) • Document timeline 	N.A.	N.A.
TDTlighthouse	<ul style="list-style-type: none"> • List of topics (important terms) 	<ul style="list-style-type: none"> • Cluster visualisation 	N.A.
TimeMine	<ul style="list-style-type: none"> • Topic timeline 	N.A.	<ul style="list-style-type: none"> • List of important terms
Topic Tracking Visualisation	<ul style="list-style-type: none"> • List of topics • Document content 	<ul style="list-style-type: none"> • Boxes visualisation 	<ul style="list-style-type: none"> • List of important terms

*N.A.: not available

Table 2.2: Comparison of *i*TDT features

Most of the *i*TDT interfaces reviewed have the Document View (DV) as the important component that displays information such as the document timeline, document content and the list of topics or documents. Meanwhile Cluster View (CV) is also an important component which presents the stories or documents by visualising them in a cluster or in a box form. Finally Term View (TV) displays the important terms to the user. The exploration and combination of these three views (Document, Cluster and Term View) with features such as cluster visualisation and the timeline on the user interface could be effectively used together to perform the TDT tasks. Based on the works reviewed, none of them measured the effectiveness of their approach and features, applied on the interfaces, from a formal user aspect. Most of them reported on the effectiveness of the technique to the system performance using IR and TDT style evaluation. Past research has proven that user interfaces can significantly improve the effectiveness of the TDT task (Allan et al., 2005). Therefore, the challenging questions are how to effectively analyse and present news in a meaningful and efficient manner, and what kinds of additional and critical information will contribute to an interactive TDT interface design. This will be examined thoroughly in Chapter 5.

2.6 Chapter Summary

The background and motivation behind the work are described and presented in this chapter. The identified components and features of *i*TDT have motivated me to strive for the design and implementation of user interfaces in this work. It is important to identify which features and approaches will be effective and to determine how TDT interfaces can be changed to support interactive TDT tasks. This motivates me to investigate and to evaluate whether the components and features identified may help a journalist to perform the TDT task better. In Chapter 4, I discuss Named Entity Recognition in TDT and *i*TDT and describe a small pilot study which aims to understand named entities from the journalists' perspective.

Chapter 3

Methodology

3.1 Introduction

The research methodology employed in this thesis uses the two complete cycles of a three-stage incremental design software development model (Booch, 1991). Each cycle contains the following three stages:

- a. Requirements analysis: The research is empirically grounded by exploratory studies to develop an understanding of the requirements for *i*TDT interface. A review of *i*TDT works is discussed in Chapter 2 and the results from a pilot study which is discussed in Chapter 4 provide leads to the design and approach used in *i*TDT interface.
- b. Design and prototyping: Findings from the exploratory work are used to motivate the design and implementation of an *i*TDT interface, which is called an Interactive Event Tracking System (*i*Event) interface. This is discussed in Chapter 5.

- c. Evaluation: the *iEvent* interface is evaluated using user and task based evaluation. A user experiment was conducted which involved journalists. They had to perform Tracking and the Detection tasks. The evaluation on the usability of the features introduced is discussed in Chapter 6 and the evaluation on the usefulness of Named Entity Recognition is discussed in Chapter 7.

The research methodology is depicted in Figure 3.1.

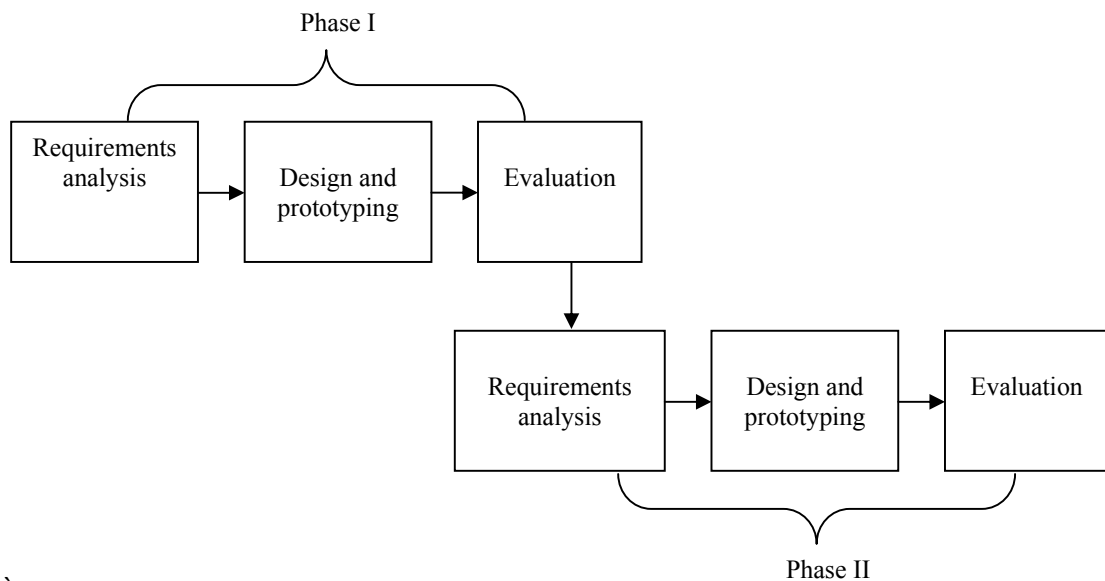


Figure 3.1: The incremental design methodology

3.2 Requirement Analysis

3.2.1 Review of Interactive TDT Systems

In Chapter 2 (Section 2.4), I have reviewed four interactive Topic Detection and Tracking (*iTDT*) systems such as the Event Organizer (Allan et al., 2005), TDTLighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000) and Topic Tracking Visualisation tool (Jones & Gabb, 2002). These were an example of TDT

works that investigated certain approaches to improving TDT system performance using GUI. I discussed the features and the approaches used and how it motivates this work. As a result, I have identified the important components and features of the *i*TDT interface. This is important in supporting journalists in performing the interactive TDT tasks.

I discovered that most of the *i*TDT interfaces reviewed have the Document View (DV) as the important component that displays information such as the document timeline, document content and the list of topics or documents. Meanwhile Cluster View (CV) is also an important component which presents the stories or documents by visualising them in a cluster or in a box form. Finally Term View (TV) displays the important terms to the user.

3.2.2 Pilot Study

A pilot study was conducted to provide me with the initial understanding of how journalists use named entities across news domain. It was aimed at understanding named entities from a journalism perspective and investigating their utility in *i*TDT. I measure named entities' distribution across news domains, the importance of named entities across news domains and the level of agreement in the keywords given by the participants. The details of this, such as the methodology and the findings, are discussed in Chapter 4 (Section 4.5).

The implications from this pilot have led to the interface design and approach. Results from the pilot study have motivated the implementation of two setups in *i*Event (named entities and keywords Setup) and a user experimental task.

3.2.3 Pilot Test

A pilot test was carried out prior to the user experiment in which four participants evaluated *i*Event and completed the user tasks. Two of the participants were research

staff, one was a research student from the Department of Computer and Information Sciences (CIS), and one was a postgraduate student from the Scottish Centre for Journalism Studies (SCJS) at the University of Strathclyde. The aim of the pilot test was to investigate the functionality of the *iEvent* interface and to receive critical feedback of the user tasks. The participants were given the same tasks used in the user experiment as reported in Section 3.4.4 (Evaluation Tasks). In the Tracking task, the participants had to track the cluster that contained the given topic and in the Detection task, they had to identify the topic dealt with by a specific cluster.

Results from the pilot test revealed that all participants agreed that they required a link from the document histogram and term histogram to point to the specific document id. They felt that if this approach was applied on the interface it would help them to perform the tasks better and save them time. 50% of the participants required document snippets that allowed them to skim over the listed document in the cluster without having them to click on each document. 50% of the participants agreed that the date format in Document View should have been written as 05/01/1998 instead of displaying CNN19980105.1130.0033 since participants were having difficulty in understanding the format. 50% of the participants also required the interface to record the term that they clicked as a tracking method to identify the cluster in Cluster View.

The pilot test allowed me to evaluate the interface and as a consequence, to resolve some interface design issues, to obtain a better understanding of participant interaction with the interface, and to obtain a better understanding of the effectiveness and functionality of the features on the *iEvent* interface. It also gives valuable feedback on how I should conduct the experiment especially the training session, the time taken for each task and minor changes to the words used in the questionnaires.

3.3 Design and Prototyping

The design and approach used in *iEvent* are based from the review of *iTDT* works which was discussed in Chapter 2 and the results from a pilot study which is discussed in Chapter 4. As a result, *iEvent* is composed of three components which are Cluster View (CV), Document View (DV) and Term View (TV). It has two settings; Setup 1 is the baseline setup that uses keywords and Setup 2 is the experimental setup that uses NER. These are discussed in greater detail in Chapter 5.

3.4 Evaluation

This section discusses details of the user experiment towards the evaluation of *iEvent*.

3.4.1 Experimental Data

The system uses a selection of 1,468 documents from TDT2 and TDT3 dataset which comes from the CNN news resource. The size of the collection is sufficient to generate clusters to be displayed on the user interface, therefore I believe that the number of documents selected is enough for the user experiment.

I used a copy of the dataset from the Event Threading Experiment of Nallapati et al. (2004). 28 topics from the TDT2 corpus and 25 topics from the TDT3 corpus were chosen. The reason for choosing only CNN news is that the stories from this source tend to be short and precise, and modelling such stories would be a useful first step before dealing with more complex datasets. Table 3.1 shows the TDT2 and TDT3 topics, including the number of documents in each topic and topic names of the selected corpus.

	Topic Id	Topic Name	# Doc
1.	20001	Asian Economic Crisis	30
2.	20002	Monica Lewinsky Case	30
3.	20012	Pope visits Cuba	30
4.	20013	1998 Winter Olympics	30
5.	20015	Current Conflict with Iraq	30
6.	20018	Bombing AL Clinic	30
7.	20019	Cable Car Crash	30
8.	20021	Tornado in Florida	30
9.	20022	Diane Zamora	23
10.	20026	Oprah Lawsuit	30
11.	20031	John Glenn	22
12.	20032	Sgt Gene McKinney	30
13.	20033	Superbowl '98	30
14.	20041	Grossberg baby murder	19
15.	20042	Asteroid Coming??	16
16.	20044	National Tobacco Settlement	30
17.	20047	Viagra Approval	30
18.	20048	Jonesboro shooting	30
19.	20056	James Earl Ray's Retrial?	27
20.	20065	Rats in Space!	30
21.	20070	India - A Nuclear Power?	30
22.	20071	Israeli-Palestinian Talks (London)	30
23.	20076	Anti-Suharto Violence	30
24.	20077	Unabomber	30
25.	20086	GM Strike	30
26.	20087	NBA finals	30
27.	20091	German Train derails	17
28.	20096	Clinton-Jiang Debate	23
29.	30002	Hurricane Mitch	30
30.	30003	Pinochet Trial	30
31.	30004	Houston Chukwu Octuplets	30
32.	30005	Osama bin Laden Indictment	23
33.	30006	NBA Labor Dispute	30
34.	30008	November APEC Summit	18
35.	30012	Leonid Meteor Shower	21
36.	30015	October Holbrooke-Milosevic Meeting	30
37.	30023	Kevorkian Trial	25
38.	30024	Gingrich Resigns	30
39.	30031	Space Shuttle Launch	30
40.	30042	PanAm Bombing Trial	29
41.	30045	Mobil-Exxon Merger	24
42.	30046	House Speaker-Elect Livingston Resigns	22
43.	30047	Space Station Module Zaria Launched	30
44.	30050	US Mid-term Elections	30
45.	30053	Clinton's Gaza Trip	30
46.	31008	Matthew Shepard Murder	30
47.	31013	Abortion Doctor Slepian Killed	30
48.	31026	Yankees vs. Padres in World Series	30
49.	31031	US Federal Budget	30
50.	31032	Yeltsin's Illness	21
51.	31033	Microsoft Anti-Trust Case	30
52.	31036	Joe DiMaggio Illness	30
53.	31038	American Embassy Bombing Trial	28
	Total		1468

Table 3.1: TDT2 and TDT3 topics

3.4.2 Evaluation Participants

The participants were a mixture of journalists and postgraduate journalism students from the Scottish Centre for Journalism Studies (SCJS), University of Strathclyde. 20 participants were recruited, of which 13 were female. 50% of the 20 were journalists and the remaining were students. The average participant age was 30-40 years. In terms of education background, 70% of participants had or were pursuing a postgraduate degree, 25% had an undergraduate degree and 5% had a Higher National Diploma⁶ (HND). 85% of participants had working experience in journalism with 30% having more than 10 years of experience and the typical journalist type was that of the daily news reporter. Participants were paid £20 for participating in the evaluation.

3.4.3 Experiment Evaluation

Text classification effectiveness is often based on two measures. It is common for Information Retrieval (IR) experiments to be evaluated in terms of precision and recall, where precision is the proportion of retrieved documents which are relevant and recall is the proportion of relevant documents retrieved.

In TDT, *system error rates* are used to evaluate text classification. These errors are system *misses* and *false alarms*, and the accuracy of a system improves when both types of errors decline. In topic tracking, *misses* occur when the system does not track the relevant documents, and *false alarm* occurs when the system labels the non-relevant documents as relevant. In new event detection, *misses* occur when the system does not detect a new event; and *false alarm* occurs when the system indicates a document contains a new event, when in truth it does not. In addition to system error rates, I report the traditional text retrieval measures of recall and precision. One prevalent approach is

⁶ A Higher National Diploma (HND) is a higher education qualification in the United Kingdom. This qualification can be used to gain entry into universities, and is considered equivalent to the first two years of a university course.

to evaluate text classification using the F1-measure (Lewis & Gale, 1994) as shown in Table 3.2, which is a combination of recall and precision from IR (Van Rijsbergen, 1979; Kowalski, 1997).

Effectiveness measures	Methods
Recall (R)	$a/(a+c)$
Precision (P)	$a/(a+b)$
F1-Measure	$2(PR)/(P+R)$
Miss Rate (M)	$c/(a+c)$
False Alarm Rate (F)	$b/(b+d)$

Table 3.2: TDT standard performance measures

The methods for calculating the effectiveness measures for new event detection, clustering and tracking are the same. These are summarised in Table 3.3.

	On topic	Not on topic
In cluster	a	b
Not in cluster	c	d

Table 3.3: Methods for calculating the effectiveness measures in TDT

The retrieved documents refer to those that have been classified by the system as positive instances of an event and the relevant documents are those that have been manually judged relevant to an event.

I treat each cluster as if it were the desired set of documents for a topic. The selection of topics for the Tracking and the clusters for the Detection task in this experiment is based on the F1-measure (Lewis & Gale, 1994; Pons-Porrata et al., 2004) as shown in Table 3.4 and Table 3.5.

Code	TopicId	Cluster	Topic	F1-measure (%)
T1	20026	10	Oprah Lawsuit	94.7
T2	20044	19	National Tobacco Settlement	58.1
T3	20019	25	Cable Car Crash	92.1
T4	20048	5	Jonesboro shooting	55.3
T5	30015	50	October Holbrooke-Milosevic Meeting	94.7
T6	30045	9	Mobil-Exxon Merger	56.0
T7	20091	46	German Train derails	91.4
T8	20042	22	Asteroid Coming??	52.6

Table 3.4: Topics in the Tracking task (F1-measure)

Code	TopicId	Cluster	Topic	F1-measure (%)
D1	31032	52	Yeltsin's Illness	77.8
D2	20022	11	Diane Zamora	77.3
D3	20096	13	Clinton-Jiang Debate	73.7
D4	20032	8	Sgt Gene McKinney	72.3
D5	30005	17	Osama bin Laden Indictment	71.4
D6	30024	49	Gingrich Resigns	50.9
D7	20018	1	Bombing AL Clinic	50.4
D8	31036	45	Joe DiMaggio Illness	49.2
D9	30002	34	Hurricane Mitch	48.3
D10	30047	23	Space Station Module Zaria Launched	44.2

Table 3.5: Clusters in the Detection task (F1-measure)

3.4.4 Evaluation Tasks

The participants were given two tasks: Tracking and Detection.

A. Tracking Task

The tracking task is defined as tracking the cluster that contains the identified topic. The participant has to track the cluster that contains the given topic and show that the system provides a sufficient amount of information on the event. This is in line with the journalist's task of reporting news. There are two sub activities in this task which are Reporting and Profiling. Reporting is defined as writing an article about a given topic. It requires the participant to write an article of a topic by drafting the important facts. Meanwhile Profiling is defined as providing the important keywords as a profile for a topic. The tasks were designed to follow naturalistic news monitoring behaviour by the participants. I wanted participants to interact with *iEvent* as if they were performing their own everyday news monitoring and reporting tasks. To do this, the tasks were placed within simulated situations (Borlund, 2000; Borlund & Ingwersen, 1997) where the given task's scenarios should reflect and promote a real news monitoring situation. Figure 3.2 shows an example of a simulated situation in a Topic Tracking task (see Appendix A.3)

SIMULATED SITUATION
<p>It is March 1998. Two students, 13 year old Mitchell Johnson and 11 year old Andrew Golden who were arrested for killing four female students, a teacher, and wounding other students at their middle school in Jonesboro, Arkansas have been put on trial. The trial caused a huge national debate due to the age of the two students. The sentence, which was announced yesterday, was that they should be treated as juveniles and remains in custody if a judge deem they're delinquent. However, a national debate about this trial and whether or not to try for the death penalty is still ongoing.</p> <p><i>You have been asked to write an article on the outcome of the trial.</i></p> <p>Topics likely to be important are the shooting; the custody and trial; the debate and reaction to the shootings; and the school situation after the tragedy.</p>

Figure 3.2: Simulated situation

The process to perform the Tracking task is that:

- i. The participants receive a topic summary.
- ii. The participants are asked to track the related cluster based on the information provided in the topic summary.
- iii. Once they have identified the related cluster, they will investigate the documents in it.
- iv. They are then asked to perform the Reporting task by drafting the important facts or points.
- v. Next they will list out the identified cluster.
- vi. They are then asked to perform the Profiling task by creating a profile of useful keywords for that topic.
- vii. Finally they complete a questionnaire about their opinion on the features of *i*Event during the Tracking task.

B. Detection Task

The detection task is defined as identifying the topic dealt by a specific cluster. This is in line with the journalist's task of identifying some important events that happened on a specific day. Figure 3.3 shows an example of Topic Detection task (see Appendix A.4).

Please indicate the topic that Cluster 52 is dealing with. If you think there is more than one topic in this cluster, please rank maximum 3 topics in the list below:

	Topic Name	Rank
1	Yeltsin's Illness	
2	Gingrich Resigns	
...		
...		
...		
20	Clinton's Gaza Trip	

Figure 3.3: Topic Detection task

The process to perform the Detection task is that:

- i. The participants receive a specific cluster.
- ii. The participants are asked to detect the topics from the documents contained in a specific cluster using any features of *iEvent* to perform this task.
- iii. They are then asked to give a ranking if they felt that the specific cluster contained more than one topic from the list of twenty topics given.
- iv. Finally they completed a questionnaire about their opinion on the features of *iEvent* during the Detection task.

There were eight topics for the Tracking task and four clusters for the Detection task in two sessions. After completion of the tasks, participants completed a questionnaire about using the interface. They were given two hours to attempt the entire Tracking task and were given 15 minutes for each topic. While in the Detection task, they had 40 minutes to complete it and were given 10 minutes for each cluster. The whole user experiment took about 2 hours 40 minutes to 3 hours excluding a short training session. The time assigned to each task was sufficient based on the feedback received from the Pilot Test as reported in Section 3.2.3 (Pilot Test).

Participants had a chance to perform the tasks using the interface. A Latin square design (Spärck-Jones, 1981; Spärck-Jones & Willet, 1997; Doyle, 1975) is used and the experimental design is pictured as in Figure 3.4. It allows me to evaluate the same topic using different setups. The order of topics assigned in the Tracking tasks and the order of clusters given in the Detection task were rotated to avoid any learning and fatigue factor. Topic 1 (Oprah Lawsuit) for example has a chance to be the first, second, third and fourth in order, during the Tracking task. The clusters assigned in the Detection task were invisible when participants performed the Tracking task to avoid any intersection of clusters. This is important because the intersection will affect the participants' performance because they might have come across the clusters used in the Detection

task during the Tracking task. Therefore it will make the tasks challenging to the participant.

PARTICIPANTS	SESSION 1						SESSION 2					
	TRACKING				DETECTION		TRACKING				DETECTION	
	T1	T2	T3	T4	D1	D2	T5	T6	T7	T8	D3	D4
1-10	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2
11-20	S2	S2	S2	S2	S2	S2	S1	S1	S1	S1	S1	S1

*S1=Setup1 (baseline setup); S2=Setup2 (experimental setup)

Figure 3.4: Experimental design

The selection of topics and clusters given in the user experiment has a combination of good and poor clustering performance based on the F1-measure. This is important to justify whether the *iEvent* interface helps the participant to perform the TDT tasks even though they were given a bad cluster to track or a bad topic to detect. The clustering is done using Single Pass Clustering with the threshold value, $t=1.48$ which resulted in 57 clusters. Table 3.6 shows the major and minor clusters for topics in the Tracking task. Major cluster refers to the cluster that contains most of the documents related to a topic and a minor cluster contains a smaller number of documents related to a topic. Cluster 10 is the major cluster about the Oprah Lawsuit topic since most of the documents in this topic were clustered together, while the minor cluster is Cluster 32.

Topics	Major cluster	Minor cluster
Oprah Lawsuit	10	32
National Tobacco Settlement	19	6
Cable Car Crash	25	46
Jonesboro shooting	5	35
October Holbrooke-Milosevic Meeting	50	39
Mobil-Exxon Merger	9	4
German Train derails	46	25
Asteroid Coming??	22	21

Table 3.6: Topics in the Tracking task

Table 3.7 shows the major and the minor topics for clusters in the Detection task. *Major topic* refers to the topic mostly discussed in a cluster and *minor topic* refers to the least discussed topic in a cluster. For example, Cluster 1 contains two topics where most of the documents discussed *Bombing AL Clinic* and few documents discussed the *Microsoft Anti-Trust Case* topic.

Cluster	Major topic	Minor topics
1	Bombing AL Clinic	Microsoft Anti-Trust Case
8	Sgt Gene McKinney	Grossberg baby murder Monica Lewinsky Case
11	Diane Zamora	-
13	Clinton-Jiang Debate	Yeltsin's Illness Gingrich Resigns
17	Osama bin Laden Indictment	American Embassy Bombing Trial
23	Space Station Module Zaria Launched	Rats in Space!
34	Hurricane Mitch	-
45	Joe DiMaggio Illness	NBA finals Yankees vs. Padres in World Series
49	Gingrich Resigns	Monica Lewinsky Case House Speaker-Elect Livingston Resigns
52	Yeltsin's Illness	-

Table 3.7: Clusters in the Detection task

3.4.5 Evaluation Procedure

Each participant was asked to attempt each of the tasks they had been given. The order in which the tasks were assigned and the setup used for each task was determined by the experimental design shown in Figure 3.4 (Section 3.4.4). The order of the setups assigned in the Tracking tasks and the order of clusters given in the Detection task were rotated to avoid any learning factor. This principle also applied during the Training Session (Section 3.4.6).

Each participant session lasted between one and one-and-a-half hours, depending on the time taken to complete the assigned tasks and the time taken by the participant to complete the questionnaires. Participants were offered a short break (5 to 15 minutes) after the first session. Each session consisted of the following steps:

1. Participants were welcomed and asked to read the introduction to the experiment provided on an 'Information Sheet' (Appendix A.1). This set of instructions was developed to ensure that each participant received precisely the same information. Participants could retain the information sheet after the experiment.
2. The participants were given a short overview of what the experiment would entail. I also explained my role in this experiment i.e. to observe participant interaction with the systems, to provide participants with technical support and to remind participants on the time taken in performing the tasks.
3. Participants were then asked to complete an *Entry* questionnaire (Appendix A.2). This provided background information on the participant's education, work experience and previous experience of news network tools used.
4. Participants were given a demonstration of the *iEvent* interface with both setups by following the experimental design as shown in Figure 3.4. This includes the features available on the interface, followed by a training session. The training session was the same for all participants using both setups. It gave participants a chance to familiarise themselves with the interface. Participants could ask questions or ask for general assistance at any time during the session.
5. Tracking task (Appendix A.3)
 - a. Once comfortable with *iEvent*, participants were asked to perform the Tracking task. There are two sub-activities in this task which are Reporting and Profiling. Reporting requires the participant to write an article on a topic by drafting the important facts while in Profiling: the participant had to make a profile of a story by providing important keywords. They were given 15 minutes to search and could stop early if they were unable to find any more relevant information.

Searching in this experiment refers to identifying the cluster related to a given topic.

- b. After completing the search (successfully or otherwise), the participant was asked to complete the questionnaire.
 - c. The remaining tasks were given to the participant in the second session using a different setup, following steps 5a-b. Participants were offered a short break after the first session.
6. Detection task (Appendix A.4)
- a. Participants were given 10 minutes to search and could stop early if they were unable to find any more relevant information. Searching in this experiment refers to detecting the topic for a given cluster.
 - b. After completing the search (successfully or otherwise), the participant was asked to complete the questionnaire.
 - c. The remaining tasks were given to the participant in the second session using a different setup, following steps 6a-b.
7. At the end of the experiment, participants were asked to complete the post-evaluation questionnaire and an informal post-experiment interview was conducted (Appendix A.6). The post-evaluation questionnaire compares participants' performance between setups and therefore the findings are discussed in Chapter 6.

The start time to perform the task was defined as the moment where the participant started using *iEvent* and 15 minutes later is defined as the end time to perform the Tracking task for a topic. This principle also applies to the Detection task with 10 minutes allowed for each cluster.

3.4.6 Training Session

Since *iEvent* was new and unfamiliar to participants, they received a training session on how to use it. A short time, around 30 minutes, was allocated for training at the start of

the experiment. In all cases this appeared sufficient for participants to familiarise themselves with *iEvent*. The training session was broken down into a series of stages:

1. I explained the purpose of *iEvent* i.e., to cluster news stories into the same group of events or topics by visualising the clusters.
2. Participants were introduced to the interface components and features that appeared in *iEvent* interface (e.g. Cluster View, Document View and Term View). I also printed a screenshot of the interface to describe the components and features of *iEvent* which I believe would help the participants to understand how *iEvent* works and perform the task better.
3. I gave participants a live demonstration of each setup using the same topic *General Motors Strike* for the Tracking task and Cluster 24 (*Pope visits Cuba*) for the Detection task.
4. A training session was issued and participants were given the chance to attempt the Tracking and the Detection tasks. It gave participants an opportunity to use *iEvent* in a realistic news tracking and detection context and become accustomed to the interface features.
5. The training session stopped once participants felt comfortable using *iEvent*.
6. Participants were allowed to comment or ask questions at any point during the session.

3.4.7 Questionnaires

Participants were asked to evaluate the interface using questionnaires. This was the main method used to elicit their opinion during the experiment.

Three questionnaires were developed and distributed to the participants at various points in the task: *Entry*, *Tracking* and *Detection* and *Post*. These questionnaires contained three styles of question; Likert scales, semantic differentials and open-ended questions. In this section each style is explained and examples provided.

A. Likert Scales

The Likert scales technique presents a set of attitude statements (Babbie, 2005). Participants were asked to express agreement or disagreement on a five-point scale. A five-point scale was preferred to seven or nine point scales as it made the analysis of participant opinion simpler and allowed trends in the results to be more easily identified. Each degree of agreement is given a numerical value from one to five where a higher value corresponds to more familiarity as shown in Figure 3.5. A total numerical value can be calculated from all the responses received.

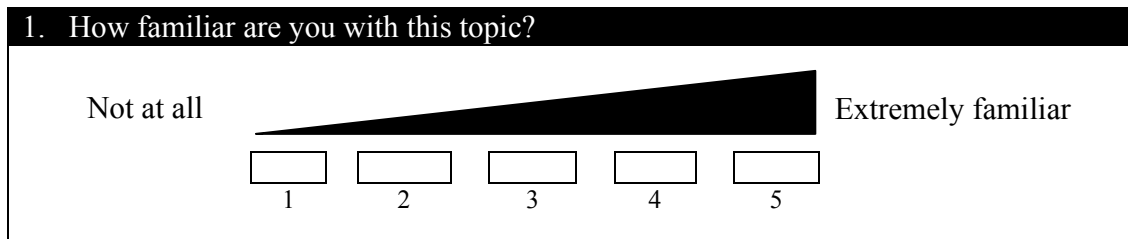


Figure 3.5: Likert scale taken from the *Tracking* questionnaire

B. Semantic Differentials

Another type of structured question is one that provides pairs of antonyms and synonyms, together with five-step rating scales. The word pairs refer to an attitude object on each continuum between the most positive and negative terms. This type of scale is called a *semantic differential* (Osgood, Suci & Tannenbaum, 1957).

2. Using this interface to track the topic was GENERALLY:

	1	2	3	4	5	
Difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Easy
Stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Relaxing
Complex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Simple
Frustrating	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Satisfying
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

Figure 3.6: Semantic differentials

Figure 3.6 exemplifies a set of five semantic differentials. Participants were asked to rate on semantic differentials for example whether the interface was difficult or easy; difficult (scale 1), average (scale 3) or easy (scale 5).

C. Unstructured Questions

In unstructured questions participants were given the chance to freely comment and reply; these questions can be described as *open-ended*. They are useful for revealing reasons why participants feel the way they do and giving them a chance to comment freely on the system, the task or the experiment in general.

3.4.8 System Logging

Log files were named based on the participant's unique identifier, the setup and topic attempted. The log file contains a header, which is written before any interaction takes place. This contained the participant identifier, the setup being used, the topic being attempted and the date and time of the experiment. Prior to starting each task I created this header using a small Java application. The interface to this application is shown in

Figure 3.7 where Participant 1 was using Setup 1 and Topic 7. Their activity and interaction with the setup will be named 1S1T7. It was not important that this interface was intelligible to experimental participants as it is used only by me.

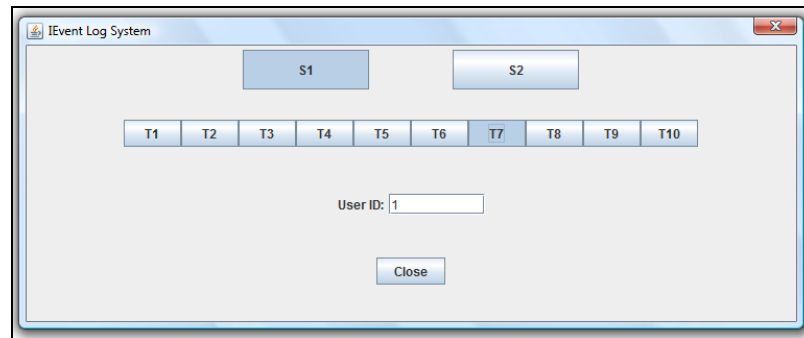


Figure 3.7: Log system

All participant interaction with iEvent was also logged as a ‘<instance><timestamp>’ pair and the timestamp was written as the number of milliseconds elapsed from midnight, January 1, 1970. This is a Java default and it allowed times to be easily parsed and compared. I constructed the output stream to ensure the instances that have been labelled with a specific tag will be printed. Table 3.8 shows the details of the tags used to denote the instances from the log files.

Tag	Meaning
Setup [#]	Setup number
Cluster [#]	Cluster number
ClusterDescB	Cluster Description Button clicked
Doc [Cluster #][DocId #]	Document viewed
Page [Cluster #][Page #]	Page viewed
TermDoc [#]	Term in document content clicked
TermDropList [Cluster #][Term #]	Term in Drop Down List clicked
TermTV [#]	Term in Term View clicked

Table 3.8: Interaction tag

3.5 Chapter Summary

This chapter summarised the research process and described how each methodological component such as the results from the Pilot Study, contributed to the design and implementation of the interface. The user tasks were designed to support journalists in performing the TDT tasks in a way attuned with their task, therefore I believe the methodology in this work is appropriate for *i*TDT evaluation.

Chapter 4

Pilot Study on the use of Named Entity Recognition for *i*TDT

4.1 Introduction

The literature reviewed in Chapter 2 revealed that there are interactive Topic Detection and Tracking (*i*TDT) interfaces which have been designed to assist users but none have been properly evaluated. In Chapter 2, I identified the useful components and features of *i*TDT while in Chapter 4 I review Named Entity Recognition (NER) in *i*TDT. This also leaves the question of how useful NER is when applied to *i*TDT in allowing users to perform Topic Detection and Tracking (TDT) tasks. Named entities provide important and significant information to journalists when writing news articles (Dijk, 1983). Therefore it is important to understand the usefulness of named entities when the journalists perform a searching task.

In this chapter I discuss Named Entity Recognition (NER) in TDT and *i*TDT. Then I explain the use of named entities from the perspective of journalism and finally I discuss

a small pilot study to provide an initial understanding of the use of named entities among the journalists. The focus of this chapter is the review of related works on NER in *iTDT* and how they have motivated me to conduct the pilot study. I also relate the findings of the pilot study to the user interface design and the user task. This motivates the use of named entities to create context in the interface. I also address the motivation to provide the user with two settings i.e. the keywords setup and named entities setup for evaluation at the end of this chapter. The keywords are sometimes referred to as the *bag of words* because there is no significant order and they do not have a specific semantic or meaning.

4.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) is widely used in Information Extraction (IE), Question Answering (QA) or other Natural Language Processing (NLP) applications. It was first used in the Message Understanding Conferences (MUC) which influenced IE research in the U.S. in the 1990's (Grishman and Sundheim, 1996). At the sixth conference (MUC-6) in 1995 the task of NER and co-reference were added. At that time, MUC focused on IE tasks where structured information of company activities and military related activities was extracted from unstructured text, such as newspaper articles. In the course of systems development, people noticed that it was important to recognise information units such as names, including personal, organisational and location names, and numeric expressions including time, date, money and percentage expressions. Extracting these entities was recognised as one of the important sub-tasks of IE. As this task is relatively independent, it has been evaluated separately in several different languages, e.g. Japanese, Chinese and Spanish in Multilingual Entity Tracking (MET) project. Outside of the U.S., there have been several evaluation-based projects for named entities, as one of the tasks of Information Retrieval and Extraction Exercise (IREX) in Japan (Sekine & Isahara, 2000). There has also been evaluation for named entities as in the shared task in the Conference on Computational Natural Language Learning (CoNLL) in 2002 and 2003 for four languages, English, German, Dutch and

Spanish. In the IREX project, a new category known as *artefact*, an example of which might be *Odyssey* as a book title or *Windows* as a product name, was added to the original MUC categories. The named entities task in MUC was inherited by the Automatic Content Extraction (ACE) project in the U.S., where two new categories were added; Geographical and Political Entities (GPE), such as *France* or *New York*; and Facility, such as *Empire State Building*.

There are situations where wider IE is required for specific scenarios, like *rocket launch* or *disease outbreak*, which require *the names of rockets* or *disease names*. The wider the IE tasks become, the more categories of named entities are needed. QA research aims to make a system which can produce an answer like *The Turing Award* to a question like ‘*What is the name of an international prize in computing that Barbara Liskov received in 2008?*’. NER plays an important role in creating such a system. Typical systems analyse the answer type from the question sentence, such as, from *prize*, and the system searches for an answer of the analysed type based on evidence such as a keyword in near context. There is an urge for a large number of categories in order to create a system capable of answering a wide variety of questions. Also, there are new fields where the named entities related task becomes an important component technology. For example, in bioinformatics, recognising names of proteins or genes is crucial. As a result, there are on-going efforts to make extended named entities (Sekine, Sudo & Nobata, 2002).

The importance of NER was highlighted by the OKKAM⁷ European Project. It conducted a survey to investigate how people describe entities such as persons, organisations, locations, events and artefacts. This study investigates some aspects of the use of keywords in Web searching. The OKKAM project aims at enabling the Web of Entities, namely a virtual space where any collection of data and information about any type of entities published on the Web can be integrated into a single virtual, decentralised and open knowledge base (Bouquet et al., 2008). Therefore, OKKAM will provide a scalable and sustainable infrastructure, called the Entity Name System (ENS),

⁷ <http://fp7.okkam.org>

available to content creators, editors and developers. This will support them to easily find public identifiers for the entities named in their contents or services, use them for creating annotations, and build new network-based services which make essential use of these identifiers in an open environment like the Web.

This section has demonstrated that NER is getting more attention in Information Extraction (IE), Question Answering (QA) and Natural Language Processing (NLP). In the next section I discuss named entities from a journalism perspective and the *state of the art* of NER in interactive Topic Detection and Tracking (*iTDT*).

4.3 Named Entities from a Journalism Perspective

Journalists are trained to apply *nine principles in journalism* (Kovach & Rosenstiel, 2001; Stovall, 2004) for ethical conduct in order to provide society with accurate and reliable information. They are:

1. Journalism's first obligation is to the truth

Journalists should be as transparent as possible about sources and methods so people can make their own assessment of the information. Due to information overload, people need identifiable sources dedicated to verifying that information and putting it in context.

2. Its first loyalty is to citizens

While news organizations answer to many constituencies, including advertisers and shareholders, the journalists in those organizations must maintain commitment to society and the larger public interest above any other if they are to provide the news without fear or favour. This commitment to the society first is the basis of a news organization's credibility.

3. Its essence is a discipline of verification

Journalists rely on professional discipline for verifying information. When the concept of objectivity originally evolved, it did not imply that journalists are free of bias. It called, rather, for a consistent method of testing information, a transparent approach to evidence so that personal and cultural biases would not undermine the accuracy of their work.

4. Its practitioners must maintain an independence from those they cover.

Independence is an underlying requirement of journalism and the basis of reliability. Independence of spirit and mind is the principle journalists must keep in focus.

5. It must serve as an independent monitor of power

Journalism has an unusual capacity to serve as watchdog over those whose power and position most affects the society. As journalists, they have an obligation to protect this watchdog's freedom by not exploiting it for commercial gain.

6. It must provide a forum for public criticism and compromise

The news media are the common carriers of public discussion, and this responsibility forms a basis for the journalist's special privileges. This discussion serves society best when it is informed by facts rather than prejudice and belief.

7. It must strive to make the significant interesting and relevant

Journalism is storytelling with a purpose. It should do more than gather an audience or catalogue the important. This means journalists must continually ask what information has most value to the society and in what form.

8. It must keep the news comprehensive and proportional

Keeping news in proportion and not leaving important things out are also the basis of truthfulness. Journalists create a map for people to navigate society. Inflating events for sensation, neglecting others, stereotyping or being disproportionately negative all

make a less reliable map. Therefore journalists have to follow the *Five W's* and the *H* guideline; *Who, What, When, Where, Why* and *How* when writing news stories. In order to keep stories new to the reader, journalists are encouraged to observe and express facts about newsworthy events in text. Thus, journalists often apply the following essentials in constructing a news story:

- a. **Who** (the subject in the story): Who is involved? Who made a scientific discovery? Who is speaking at a forum? Who organised the show?
 - b. **What** (the action that prompted the story): What is the nature of the news story or event? Is it a scientific discovery, a student activity, an appointment to a professorship, an award, a talk given at a conference?
 - c. **Where** (the physical context): Where is the news or event taking place? Is it a seminar in a common room, a talk in a lecture hall, a demonstration in the Muir Lab?
 - d. **When** (the time context): When will (or did) the event take place? What time and date is the event, or when will someone be available for a user experiment?
 - e. **Why** (authoritative comments): Why is the story newsworthy? Tell readers why they should care. Who will be affected by this news and how?
9. Its practitioners must be allowed to exercise their personal conscience
Every journalist must have a personal sense of ethics and responsibility. News organizations do well to nurture this independence by encouraging individuals to speak their minds.

The guidelines in constructing a news story mentioned are one of the principles in journalism, and are in line with the criteria used in TDT to identify an event. For example, in the Asian Tsunami story (December 2004), some important sub-events were the initial destruction of the tsunami, the relief effort, and the investigation into why there were few forewarnings of the disaster. While some facts surrounding the story did not change such as *Where did the tsunami first hit?*, others changed with time such as *How many people have been confirmed dead?*. Therefore, in order to build an interface

that will assist the journalists in finding information that helps them to fully understand a story or situation, the interface must be able to handle information on *Who, What, Where* and *When*.

I also look at the discourse analysis aspect as it studies the information flow in a press article as shown in Figure 4.1. Within the news domain, discourse analysis deals with the formation of a complete news report (mainly for news in the press), while broadcast news is usually released in shorter pieces and the context is often assumed to be available for the audience. Discourse analysis is a general term that includes many approaches to analysing the use of languages, and one important application of it is the news (Dijk, 1983).

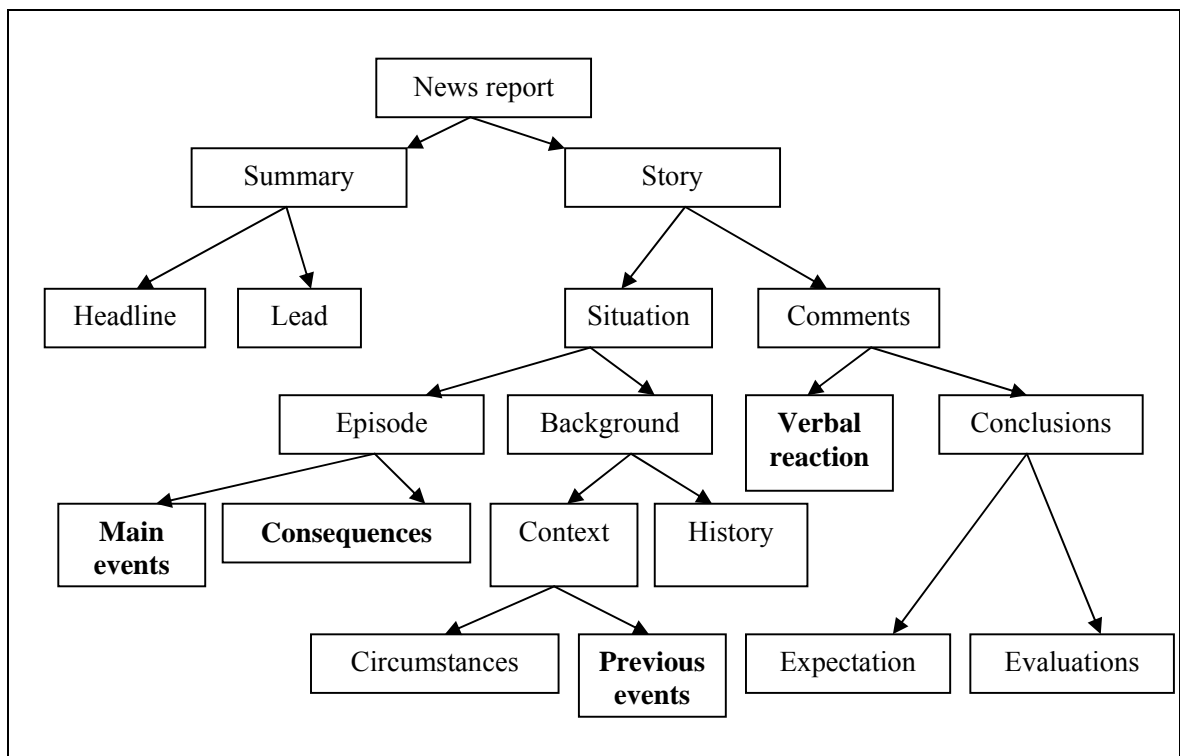


Figure 4.1: Hypothetical structure for a News Schema (Dijk, 1983).

Although assigning an event in each of the nodes, as in Figure 4.1, into a certain location of the structure is difficult, it is clear that many items can be used to visualise the structure in the form of a story timeline. A related TDT work on discourse analysis has been conducted by Nallapati et al. (2004). They implemented models of discourse analysis and investigated news schema in discourse analysis in their work on incident threading. In a similar way to discourse analysis, incident threading focuses on the contextual information in news reports. Many concepts and terms in discourse analysis apply to incident threading, mainly for the relation types among incidents. The process of identifying the incidents and generating the network is called *incident threading*. Evaluation proved that there is a significant improvement on the TDT system performance in establishing the links between related incidents. However there is no previous attempt to evaluate discourse analysis from the users' aspect.

Humans have their own ways of comprehending news information, but there are some common rules that most would follow to make news more memorable. The researcher believes that in order for a news monitoring system to facilitate users effectively in their monitoring process, it is recommended that the system should have the ability to group news according to the main topic discussed. Since human beings have reasoning abilities, they do not treat news events as isolated facts. Instead, they tend to compare new information to memory and insert it into the existing fact network, at a location next to the relevant pieces. It would be ideal if the system has the same ability to link related events, because people are very likely to be interested in both (or neither). In addition, tracing back from the new information can be a good reminder for users about things that they have already forgotten.

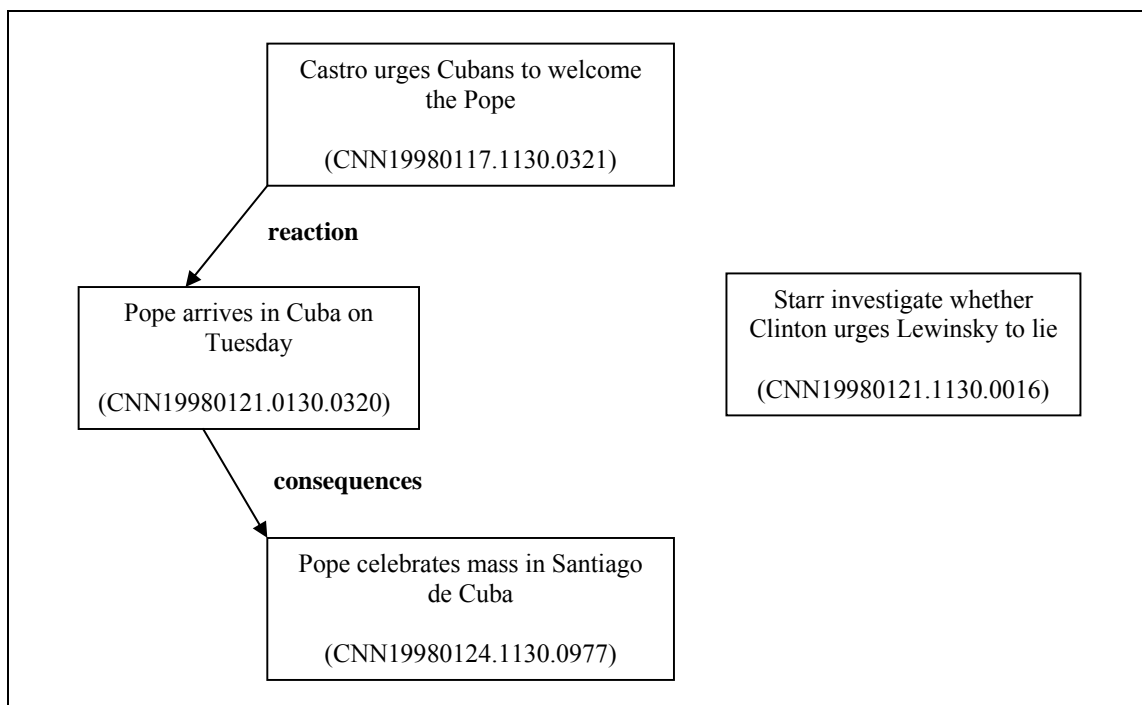


Figure 4.2: News report for topic *Pope Visit Cuba* from CNN

Figure 4.2 shows summaries of four news reports from CNN with the document identifiers from the corpus used in this work. Three of them are from the same news topic *Pope visits Cuba* and the other one is about the well-known *Monica Lewinsky* case. An ideal news organisation should place or cluster the three related reports together and show their contextual link, leaving the irrelevant information aside. This has motivated me to present and visualise discourse analysis as one of the features in an interactive TDT user interface.

4.4 Named Entity Recognition (NER) in Interactive TDT (*iTDT*)

This section discusses the related work on NER in *iTDT* such as works on document representation and user interfaces. A TDT system draws a distinction between events

and topics. For instance, during the *2007 Scottish Parliament Election*, the election is the seminal event that triggers the topic, and other stories on the same topic would be those discussing the campaign, the results, the party involved and the candidates. Events reveal specific information such as *Who, What, Where and When*.

Recent research in TDT has investigated named entities rather than keywords because TDT investigates the organisation of information by event rather than by subject (Allan et al., 1999; Kumaran and Allan, 2004; Li, 2006; Li and Croft, 2005; Makkonen et al., 2004; Otterbacher et al., 2005). Exploiting NER too has improved the accuracy of the New Event Detection (NED) systems (Makkonen et al., 2004; Kumaran and Allan, 2004; Yang et al., 2002). A good NED system would be one that correctly identifies the article that reports the terrorist attack as the first story while other stories discussing the reactions from different parts of the world, death toll, scientific discussions and the rescue efforts are considered as the other stories that make up the topic. Thus a good understanding of NER is important for a good TDT system.

4.4.1 Document Representation

Several information organisation, access, and filtering systems can benefit from different kinds of document representations than those used in traditional Information Retrieval (IR). Document representation is one of the most common and crucial stages of an information organisation and access system. Several methods and models of document representation have been proposed based on the target application. Some of them are general enough to be applicable to almost any IR-based application. However, some tasks demand a different approach to document representation. Topic Detection and Tracking (TDT) is one such domain.

Topic Detection and Tracking (TDT) began as a technology development and evaluation program (Allan et al., 1998). TDT evaluation (Fiscus & Doddington, 2002) provides a standard set of news documents with a number of topics to be tracked and a list of

relevant documents for each topic. Researchers in this area claim that technology evaluation is the main focus of TDT and does not investigate user interface issues (Allan, 2002). In addition, TDT evaluation has been carried out traditionally in a laboratory setting, which does not involve real users and real tasks. Thus, researchers in this area have focused on developing techniques and algorithms for a better TDT performance; this is also the main activity in TREC evaluation.

In recent years, NER has been receiving more attention in TDT where several efforts have been made to exploit it for document representation, in order to improve TDT systems. Yang et al. (1999) investigated and focused on *location* as a named entity for document representation. The DOREMI research group also looked at *people* and *location* named entities to obtain a final confidence score for each story (Makkonen et al., 2004). Kumaran and Allan (2004) split document representation into two parts: named entities and non-named entities. It was found that some classes of news could achieve better performance using named entities representation such as Elections, Accidents, Violence and War, New Laws, Sports News, and Political and Diplomatic Meetings. For example, the names of election candidates (Person name) are very important for stories of election class; the locations (Location name) where accidents happened are important for stories of accident class. While some other classes of news such as Natural Disasters, Criminal cases, Scandals/Hearings and Science could achieve better performance using non-named entities representation. Kuo et al. (2007) investigated the average correlation between Part-of-Speech (POS) and news genre to model New Event Detection (NED) model. They revealed that terms of different types (Noun, Verb or Person name) have different effects for different genre of stories in determining whether two stories are on the same topic. For example, the names of election candidates (Person name) are very important for stories of election class; the locations (Location name) where accidents happened are important for stories of accident class.

4.4.2 User Interface

The review of the state of the art of interactive TDT shows that NER has been used in document representation but few have applied it on user interfaces for TDT tasks. The works reviewed in Chapter 2 such as Event Organizer (Allan et.al, 2005), TDTLighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000) and Topic Tracking Visualisation tool (Jones & Gabb, 2002) have provided me with the important components and features of *i*TDT interface.

However, only Event Organizer (Allan et.al, 2005) has applied NER in their work. The clusters were labelled using important terms which consist of noun phrases or named entities. Works such as TDTLighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000) and Topic Tracking Visualisation tool (Jones & Gabb, 2002) only display the important terms instead of named entities. I believe the use of NER gives a better understanding of the news by highlighting the significant information on the *Who*, *Where* and *When* and thus users are able to understand news in a meaningful and efficient manner. Named entities and terms produce interesting information; named entities are of higher quality, but terms are more descriptive. I believe both should be used. There has been very little exploration and proper evaluation of how named entities, along with features such as cluster visualisation and the timeline on *i*TDT, might be effectively used together to perform TDT tasks.

The reviewed works of NER in TDT and *i*TDT motivated me to conduct a pilot study that aims to prove that this approach could be used effectively to create context in the interface.

4.5 Pilot Study

It is important to have a good understanding of user and information context since it influences system design (Crestani & Ruthven, 2005; Ingwersen & Järvelin, 2005; Ruthven et al., 2006; Ruthven, 1996). Thus a pilot study was performed to give an initial understanding of how journalists have categorised keywords into named entities, giving an idea of named entities distribution across news domain. This is achieved by measuring named entities distribution across different news domains. I made the assumption that '*Who*' has the highest frequency in Entertainment in comparison to Government. This is because it was mostly the name of the celebrity that was highlighted in Entertainment articles. In contrast *When* has the highest frequency in Economy since the activities in economy are concerned with timelines.

The objectives of this pilot study were to:

- a. Identify named entities distributions (*Who*, *Where*, *When* and *What*) across news domains (Politics, Economy, Government and Entertainment). These are the common news domains that provide me with the variation in identifying the distributions.
- b. Measure the importance of named entities across news domains;
- c. Measure the level of agreement in the keywords given by the participants.

These objectives guided the design of the user interface by investigating named entities from the occurrence and the importance aspects. This leads one to ask, are some types of named entities more important and more frequent in some news domains than in others?

4.6 Methodology

This section explains the corpus used and the methodology of conducting the pilot study.

4.6.1 Instruction and Online Survey

An online survey was placed on a server in the Department of Computer and Information Sciences (CIS) at University of Strathclyde as shown in Figure 4.3. It was designed using HTML and CGI scripting as a front end to receive the data via email.

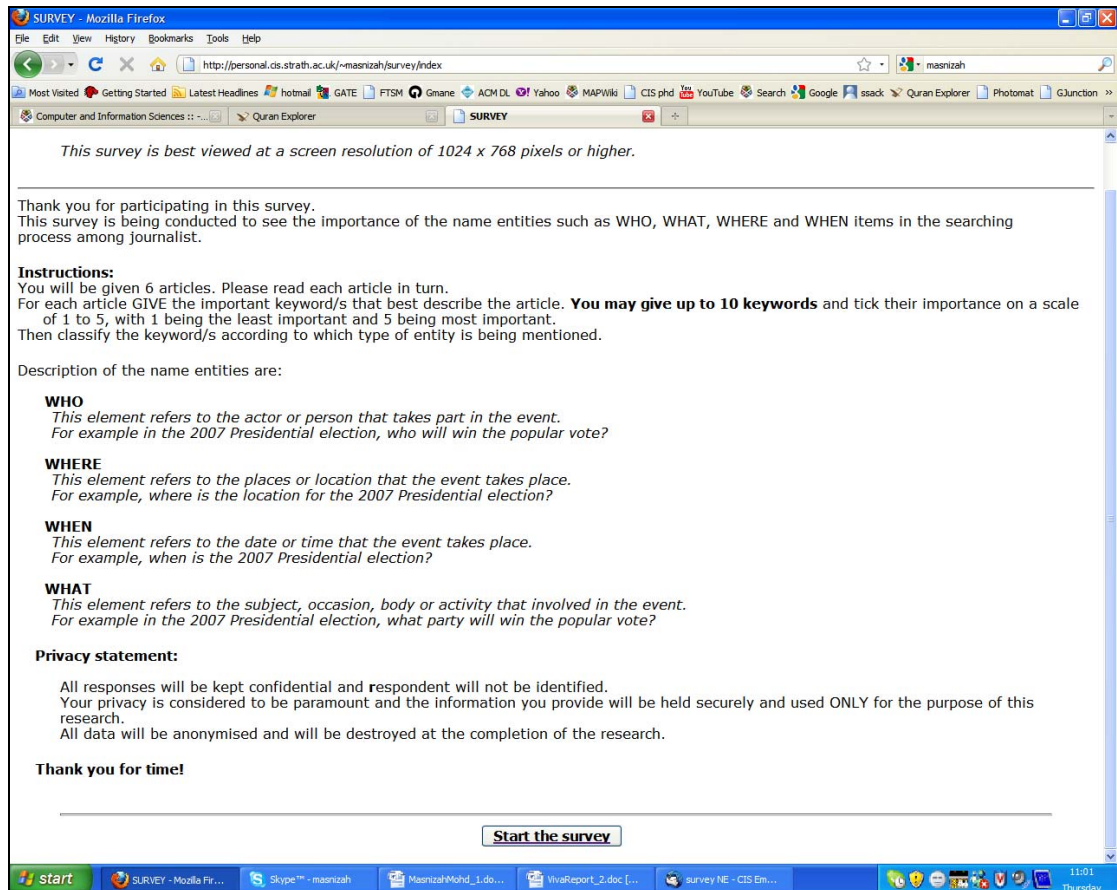


Figure 4.3: Online survey main interface

The survey aimed to identify the use of named entities such as WHO, WHAT, WHERE and WHEN in the searching process of journalists. Participants were given the privacy statement at the main page and the description of named entities such as:

- a. WHO: This element refers to the actor or person that takes part in the event.
For example in the 2007 Presidential election, who will win the popular vote?
- b. WHERE: This element refers to the places or location that the event takes place.
For example, where is the location for the 2007 Presidential election?
- c. WHEN: This element refers to the date or time that the event takes place.
For example, when is the 2007 Presidential election?
- d. WHAT: This element refers to the subject, occasion, body or activity involved in the event.
For example in the 2007 Presidential election, what party will win the popular vote?

They were given six documents and were asked to read each of them. They were then required to provide keywords that best described the document and to tick their importance on a scale of 1 to 5, with 1 being the least important and 5 being most important. Finally, they had to classify the keywords according to which type of named entity was being mentioned.

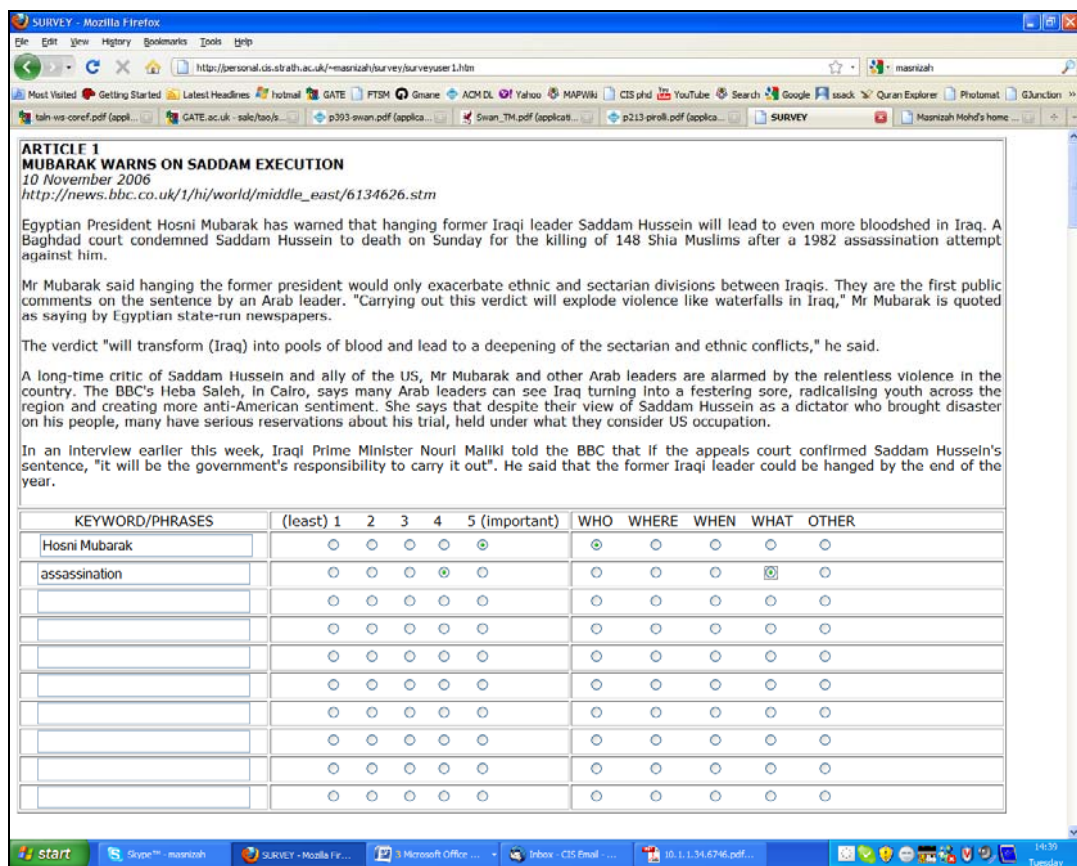


Figure 4.4: Online survey

As shown in Figure 4.4:

- Then participants were asked to read each document (e.g. Article 1 on *Mubarak Warns on Saddam Execution*).
- They were then required to provide keywords that best described the document (e.g. *Hosni Mubarak*).
- Then they were required to tick the importance of the keyword provided on a scale of 1 to 5, with 1 being the least important and 5 being most important (e.g. 5-important).
- Finally they had to classify the keywords according to which type of named entity was being mentioned (e.g. WHO).

4.6.2 Corpus and Distribution

The study consisted of two phases; the first was conducted in early February 2007, while the second took place in early March 2007. I believe the number of documents in the corpus would not have been sufficient had the study been conducted in one phase since it only involves five documents in each news domain. In addition this would not have provided me with a proper understanding on the use of named entities among journalists. 10 postgraduate students from the Scottish Centre for Journalism Studies (SCJS) University of Strathclyde were selected for this study.

Participant	Phase	Documents Distribution					
		Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
1	I	P1	E1	G1	Et1	P2	E2
	II	P6	E6	G6	Et6	P7	E7
2	I	G2	Et2	P3	E3	G3	Et3
	II	G7	Et7	P8	E8	G8	Et8
3	I	P4	E4	G4	Et4	P5	E5
	II	P9	E9	G9	Et9	P10	E10
4	I	G5	Et5	P1	E1	G1	Et1
	II	G10	Et10	P6	E6	G6	Et6
5	I	P2	E2	G2	Et2	P3	E3
	II	P7	E7	G7	Et7	P8	E8
6	I	G3	Et3	P4	E4	G4	Et4
	II	G8	Et8	P9	E9	G9	Et9
7	I	P5	E5	G5	Et5	P1	E1
	II	P10	E10	G10	Et10	P6	E6
8	I	G1	Et1	P2	E2	G2	Et2
	II	G6	Et6	P7	E7	G7	Et7
9	I	P3	E3	G3	Et3	P4	E4
	II	P8	E8	G8	Et8	P9	E9
10	I	G4	Et4	P5	E5	G5	Et5
	II	G9	Et9	P10	E10	G10	Et10

(P=Politics, E=Economy, G=Government, Et=Entertainment)

Table 4.1: Documents Distribution

The distribution of documents for the participants is listed in Table 4.1. From this distribution, every document was viewed by 3 participants such as document P1 (*Saddam verdict timing 'suspect'*) and document P6 (*Mubarak warns on Saddam execution*) were viewed by respondents number 1, 4 and 7. A total of 12 documents were given to each participant from this pilot study. The documents' distribution was based on a repeated Latin square. The reason for this was to have a balanced distribution of the documents to every participant such that each participant receives documents from the four news domains and every document will be viewed by three participants. As a result, this distribution allows the comparison of the keywords from the same document and supports the third objective of the pilot study which is to measure the level of agreement among participants.

The corpus was a collection of 40 documents chosen from CNN News, the Associated Press and Scotsman.com. Each news domain consisted of ten documents. 18% of the sources and documents were current stories related to Scotland or Glasgow as all participants were living in Glasgow. For example, documents P5 and P10 in Politics; E1, E4 and E8 in Economy; G1 in Government; and Et3 in Entertainment. These documents are shown in Table 4.2.

News domain	Doc.ID	Document
<i>Politics</i>	P1	Saddam verdict timing 'suspect'
	P2	Protests as Bush visits Indonesia
	P3	Blair meets troops in Afghanistan
	P4	Support for Labour falls to 20-year low
	P5	Nationalists say figures prove Scots subsidizing the UK
	P6	Mubarak warns on Saddam execution
	P7	Bush offers Veterans Day tribute
	P8	War on terror 'could last 30 years'
	P9	SNP 'wants to lay down law on jail sentences'
	P10	Gordon Brown has 'sold his Scottish soul' says SNP chief
<i>Economy</i>	E1	Scottish Water customers 'saving £90 a year on bills'
	E2	Rainy day for Amazon as profits drain away to £10m
	E3	Jobless total shows rise over last year
	E4	Skills gap seen as biggest threat to Scottish firms
	E5	Britain branded 'card fraud capital'
	E6	EU to call for more labour-law flexibility
	E7	Proof of Concept scheme will generate £125 million
	E8	Scottish house price rises soar above rest of UK
	E9	Finance sector now 10% of UK economy
	E10	Economic impact of immigrants in the spotlight
<i>Government</i>	G1	Students reject union membership
	G2	Poorest families may get cash for fuel
	G3	Extra money for mentors to tackle bullying
	G4	Behaviour measures target parents
	G5	PM unveils plans for 'supernannies'
	G6	Latest figures on migrants released
	G7	Dutch Plan To Ban Burqas And Masks
	G8	Low income families grants call
	G9	NHS 'needs time to balance books'
	G10	Bourne seeks Welsh spending boost
<i>Entertainment</i>	Et1	Cruise and Holmes leave Italy after wedding
	Et2	Madonna 'wanted to adopt baby girl'
	Et3	Louis: Macs Can Beat The Axe
	Et4	Virtual "Big Brother" to be launched in Second Life
	Et5	Ramsay gets his hands on Emmy award
	Et6	Tom and Katie 'had already married'
	Et7	Madonna rubbishes new adoption reports
	Et8	Penguins beat Bond at US cinemas
	Et9	X Factor's Robert upbeat on future
	Et10	OJ Interview Cancelled: Family's Relief

Table 4.2: Documents across news domains

4.7 Results

A total of 557 keywords from the 40 documents were analysed and used to identify:

- a. Named entities distributions across news domains;
- b. The importance of named entities and news domains;
- c. The level of agreement in the keywords and news domains.

4.7.1 Named Entities Distribution across News Domains

Table 4.3 summarises the distribution of named entities across the news domains. The most striking result to emerge from the data is that the dominant type of named entity was the *What* entity; its percentage of occurrence across news domains accounts for at least 40% of keywords. The highest percentage of *What* was in Economy (64.1%). However, the distribution of *When* within news domains was less than 6% and surprisingly the percentage of it was 0% in Government.

NEWS DOMAINS	<i>n</i>	NAMED ENTITIES, <i>n</i> (%)			
		<i>Who</i>	<i>Where</i>	<i>When</i>	<i>What</i>
Politics	157	57 (36.3%)	28 (17.8%)	8 (5.1%)	64 (40.8%)
Economy	128	24 (18.7%)	20 (15.6%)	2 (1.6%)	82 (64.1%)
Government	137	50 (36.5%)	13 (9.5%)	0 (0%)	74 (54.0%)
Entertainment	135	51 (37.8%)	21(15.5%)	2(1.5%)	61 (45.2%)

(Highest value shown in bold)

Table 4.3: Named entities distribution across news domains

The calculated ranges of named entities across news domains was 40%–65% for *What*, 18%–38% for *Who*, 15%-18% for *Where* and 1%-6% for *When*. This result is interesting as the *What* and *Who* are perceived as being the most frequent type of named entities

chosen in this study. *What* and *Who* are the top named entities across news domains. The sequence of named entities distribution was identical across news domains, with *What* more than *Who* followed by *Where* and the least was *When*. However results from a Kruskal-Wallis Test show there is a statistically significant difference in the distribution of named entities frequency within news domains, $\chi^2(3)=17.1$ and $p<0.001$. A Chi-square test was performed and there was strong evidence to indicate a relationship between named entities and news domains, $\chi^2(9, n=557)=32.1$ and $p<0.001$. This shows that the distribution of the type of named entity is domain dependent across Politics, Economy, Government and Entertainment.

4.7.2 The Importance of Named Entities in the News Domains

Table 4.4 summarises the distribution of the importance of named entities across news domains. Findings showed that for the Very Important level (scale 5) of named entities, *What* had the highest occurrence across news domains and the highest percentage of *What* was in Economy (63.4%). While for the Important level (scale 4) of named entities, again, *What* had the highest occurrence in Economy, Government and Entertainment except in Politics, where *Who* has a higher percentage (36.4%). For the Fairly Important level (scale 3) of named entities, *What* was the highest occurrence in Politics, Economy and Government except in Entertainment, where *Who* has a higher percentage (44.0%). Surprisingly, for the Quite Important level (scale 2) of named entities, *Where* has the highest percentage in Politics (44%) and Entertainment (100%), *What* was the highest occurrence in Economy (100%) and *Who* and *What* shared the same percentage (50%) in Government.

I used the Spearman's rank correlation test to predict the correlation between the importance of named entities and news domains. The findings indicated that there is no significant relationship ($p>0.001$) between the importance of named entities and news domains. The importance of named entities is domain independent and it shows that participants are giving different weightings on different types of named entities.

Although *What* is the top named entity, it is not necessarily Very Important across news domains.

NEWS DOMAINS	<i>n</i>	NAMED ENTITIES, <i>n</i> (%)			
		Who	Where	When	What
Very Important (Scale 5)					
Politics	77	30 (39.0%)	8 (10.4%)	5 (6.5%)	34 (44.2%)
Economy	71	17 (23.9%)	8(11.3%)	1 (1.4%)	45 (63.4%)
Government	68	30 (44.1%)	2 (2.9%)	0 (0%)	36 (52.9%)
Entertainment	78	32 (41.0%)	5 (6.4%)	1 (1.3%)	40 (51.3%)
Important (Scale 4)					
Politics	33	12 (36.4%)	8 (24.2%)	2 (6.1%)	11 (33.3%)
Economy	30	3 (10.0%)	10 (33.3%)	0 (0%)	17 (56.7%)
Government	39	10 (25.6%)	7 (17.9%)	0 (0%)	22 (56.4%)
Entertainment	30	8 (26.7%)	7 (23.3%)	1 (3.3%)	14 (46.7%)
Fairly Important (Scale 3)					
Politics	38	12 (31.6%)	8 (21.1%)	1 (2.6%)	17 (44.7%)
Economy	24	4 (16.7%)	2 (8.3%)	1 (4.2%)	17 (70.8%)
Government	26	8 (30.8%)	4 (15.4%)	0 (0%)	14 (53.8%)
Entertainment	25	11 (44.0%)	7 (28.0%)	0 (0%)	7 (28.0%)
Quite Important (Scale 2)					
Politics	9	3 (33.3%)	4 (44.4%)	0 (0%)	2 (22.2%)
Economy	3	0 (0%)	0 (0%)	0 (0%)	3 (100%)
Government	4	2 (50%)	0 (0%)	0 (0%)	2 (50%)
Entertainment	2	0 (0%)	2 (100%)	0 (0%)	0 (0%)

(Highest value shown in bold)

Table 4.4: Importance of named entities across news domains

4.7.3 Level of Agreement in the Keywords and News Domains

The level of agreement in the keywords given by the participants was calculated by using the overlap value (Voorhees and Harman, 2000) which is the intersection of

keywords divided by the union of keywords from the same document. Thus, an overlap of 1.0 means perfect agreement and an overlap of 0.0 means none of the participants agreed with the keywords given. I used the Porter stemming algorithm which is that most widely used in TDT. It works to ensure that the intersection of keywords such as launched, launching and launch were counted as a single term.

I also conflated different keywords referring to the same context and meaning to generate the synonym sets. This was generated using GATE co-reference editor. It allows co-reference chains to be displayed and edited manually. This allowed me to generate a synonym set containing words such as *unemployment* and *jobless*. Meanwhile the method for named entity co-reference resolution involves the use of the ANNIE orthomatcher , which can identify:

- *equivalent*, as defined in a synonym list: this rule is used to handle matching of names like *Nationalists* and *SNP*.
- *acronyms* like *National Health Service* and *NHS*.
- *word token match*: do all word tokens match, ignoring punctuation and word order, e.g., *Hamid Karzai* and *Karzai, Hamid*.
- *first token match*: does the first token in one name match the first token in the other, e.g., *Hamid Karzai* and *Hamid*.
- *last token match*: does the last token in one name match the other name (which must be one token only), e.g., *Hamid Karzai* and *Karzai*.
- *prepositional phrases*: matches organisation names which are inverted around a preposition, e.g., *University of Glasgow* and *Glasgow University*.

Table 4.5 shows the overlap values that ranged from 0.11 to 1.00. Four documents with perfect agreement (1.0) were *NHS needs time to balance books* (G9), *Cruise and Holmes leave Italy after wedding* (Et1), *Madonna wanted to adopt baby girl* (Et2) and *X Factor's Robert upbeat on future* (Et9). 2 documents with the lowest overlap value (0.11) were *Support for Labour falls to 20-year low* (P4) and *Bourne seeks Welsh spending boost*

(G10). The findings revealed that Entertainment has the highest value of mean overlap (0.67) compared to Politics (0.42), Economy (0.41) and Government (0.47). Finding also found disagreement in the type of named entity given to the same keyword such as in the article P9 and P2. Participants classified the keywords *Holyrood election* as *When* and *What*, and the keyword *Terrorist* as *What* and *Who*. The reason for this is not clear, but it may be due to the participant’s interest and familiarity with the genre of documents. The overlap value from this study showed the variation in the individual judgements of the keyword from the document.

Politics		Economy		Government		Entertainment	
ID	Overlap	ID	Overlap	ID	Overlap	ID	Overlap
P1	0.38	E1	0.60	G1	0.57	Et1	1.00
P2	0.50	E2	0.40	G2	0.50	Et2	1.00
P3	0.57	E3	0.57	G3	0.33	Et3	0.80
P4	0.11	E4	0.29	G4	0.80	Et4	0.60
P5	0.83	E5	0.67	G5	0.67	Et5	0.67
P6	0.33	E6	0.60	G6	0.33	Et6	0.80
P7	0.67	E7	0.33	G7	0.33	Et7	0.67
P8	0.33	E8	0.40	G8	0.33	Et8	0.43
P9	0.38	E9	0.57	G9	1.00	Et9	1.00
P10	0.50	E10	0.67	G10	0.11	Et10	0.43

(Highest value shown in bold)

Table 4.5: Overlap values for documents across news domains

4.8 Discussion

Recent research in TDT has used information on named entities and this requires more effort to understand it. Thus, this study has investigated the type and the use of named entities in news domains. The study led to a better understanding of the correlation of the type of named entities across news domains. This study revealed that there is a significant difference in the distribution of named entities within news domains. Across

news domains, *What* is the most dominant type of named entity followed by the *Who* named entity. This is interesting since the results give clues as to what named entities a participant would expect to occur more often within news domains. The study also found that there is no significant relationship between the importance of named entities and news domains, indicating that the importance of named entities is domain independent. This revealed the participants' perception on how often the named entities occurred and how important the named entities are across news domains. *When* named entity has 0% occurrence in Government but it can be a Very Important type of named entity. While the level of agreement in the keywords among participants is likely to be higher in Entertainment compared to other news domains.

The findings from this pilot study have key implications for the next step of this work in two ways:

- a. Two perceptions on the occurrence and on the importance aspect of named entities have led to the interface design and approach. I decided to provide the *Who*, *What*, *Where* and *When* in the user interface without applying any weighting on it in the document representation. The design of the user interface will have:
 - The Term View that contains information on the term frequency in a histogram form with the timeline to show its occurrence.
 - Two setups which use the keywords (baseline setup) and named entities (experimental setup) as shown in Figure 4.5.

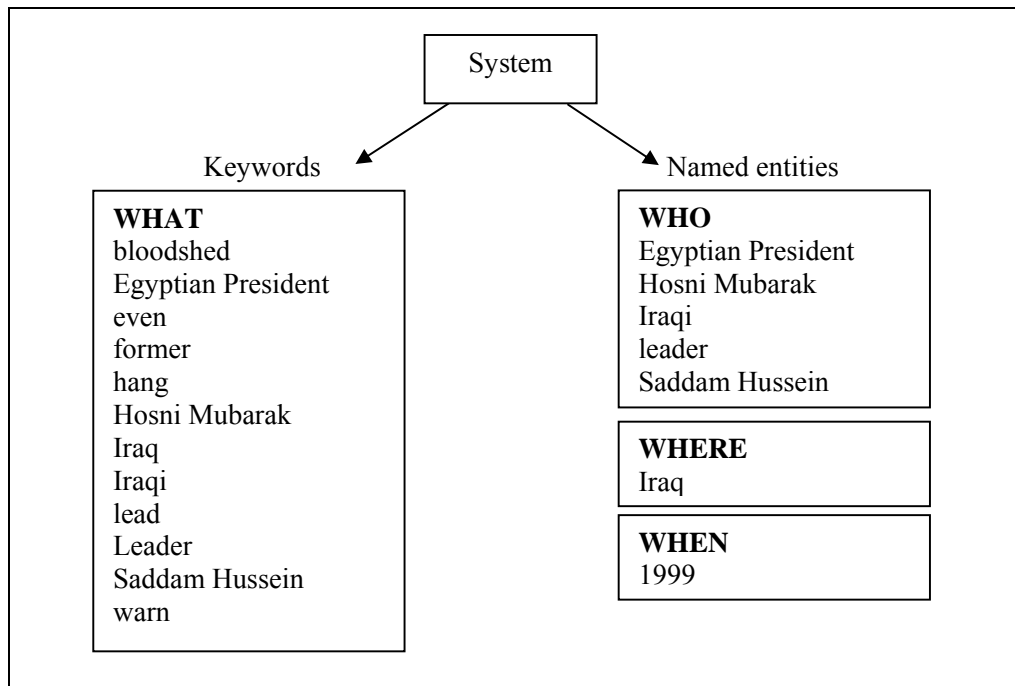


Figure 4.5: User interface approach

- b. The level of agreement has motivated the task given for the evaluation. As a result there will be a Profiling task. This is a task where the participant has to make a profile of a story by providing the important keywords. Similarly the Profiling task allows me to measure the level of agreement of the keywords given for a topic, thus I can try to relate it with which type of term being used. For example, profiling a topic on Oprah Lawsuit will give information on what are the important keywords that participants agreed most on. If all of the participants gave the keywords *Oprah Winfrey* or *Oprah*, it indicates that for this topic, named entities are mostly being used for profiling. This information is useful for measuring the effectiveness of named entities. This task was discussed in Chapter 3 (Section 3.4.4).

Very few researchers have worked on interfaces for TDT. I believe that an interactive TDT system does not only rely solely on the system performance but also on user interaction. The importance of user interaction motivated me to design and develop a

user interface for an interactive TDT system. Results from this pilot study managed to identify the context of the user interface by offering the keywords that resemble all terms and named entities that resemble the *Who*, *Where* and *When* named entities. Moreover, the level of agreement investigated in this pilot study motivates the user experimental task which is the Profiling task.

4.9 Chapter Summary

I investigate the standard keywords approach, the most used approach to TDT and the only one investigated in *i*TDT, and an approach based on NER; a novel approach in TDT and one never before evaluated in *i*TDT. NER seems to be the conventional approach used to enhance TDT system performance. Meanwhile NER has been used in document representation for interactive TDT but only Event Organizer (Allan et.al, 2005) has applied it on user interfaces for TDT tasks. One assumption of this work is that the use of NER creates the context in the interface which allows professionals, such as journalists, to perform interactive TDT tasks since it is in line with the journalism perspective. In the next chapter I will discuss the design and implementation of the user interface.

Chapter 5

***i*Event User Interface**

5.1 Introduction

Very few researchers have worked on interfaces for Topic Detection and Tracking (TDT). I believe that an interactive TDT (*i*TDT) system does not only rely on the system performance but also on the user interaction. Interfaces play a vital role in *i*TDT and I set out to design a new interface for *i*TDT that is meant to support the user in all tasks related to TDT. The importance of user interaction has motivated me to design and develop an *i*TDT interface.

In this chapter I present and describe a novel interface design that incorporates some successful features from existing TDT interfaces and that can be integrated into a single interface called Interactive Event Tracking System (*i*Event). This is a new work in TDT and the first work that both investigates the design for *i*TDT interface and evaluates it. This interface supports the user in identifying new events and tracking them in a news

stream. It also aims to cluster news stories into the groups of events or topics by visualising the clusters. It uses Named Entity Recognition (NER) as a value added component and comes with the three components; Cluster View (CV), Document View (DV) and Term View (TV).

5.2 *i*Event Architecture

*i*Event is implemented using Java Servlet version 2.5 and working under the Windows XP operating system. The *i*Event architecture is shown in Figure 5.1.

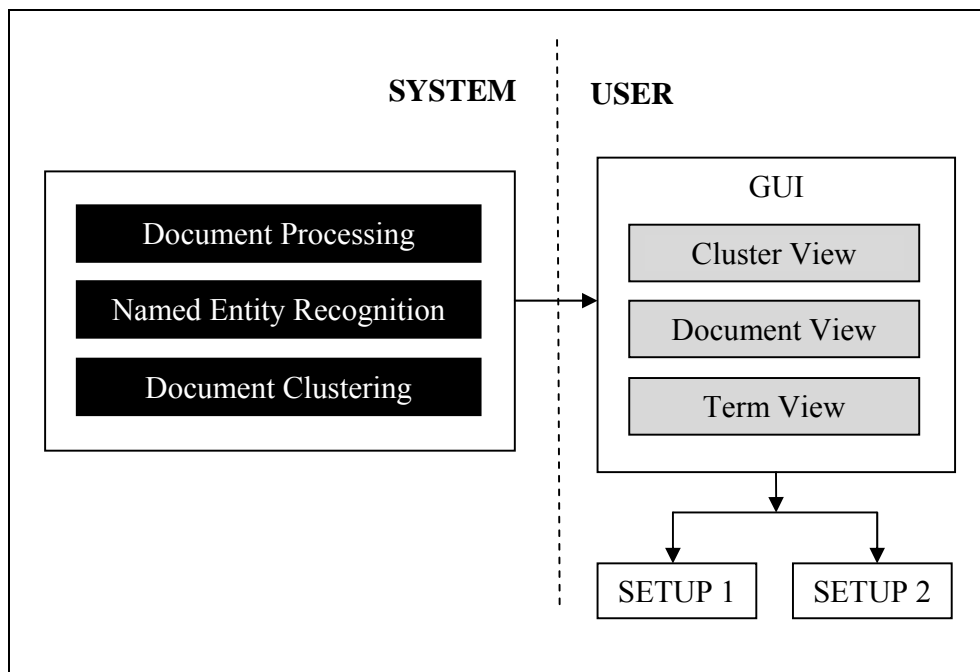


Figure 5.1: *i*Event architecture

The architecture of *i*Event consists of two sides which are the System and the User side. There are three main components in the System side; Document Processing, Named Entity Recognition (NER) and Document Clustering. There are also three components of the interface in the User side; Cluster View (CV), Document View (DV) and Term View (TV).

5.2.1 Document Processing

The system removes the tags (i.e. <DOC>, <DOCNO>, <DOCTYPE>, <DATE_TIME>, <BODY>, <TEXT>, <END_TIME>) from the documents and breaks the rest of the text into words. Then stop words⁸ are removed and Porter stemming (Spärck-Jones and Willet, 1997) is applied.

Each document was represented by a vector t_i with term as the attributes and the attribute value being its *tf.idf* weight (Salton and Buckley, 1988). This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The frequency of a term in a document (*tf*) is weighted by the inverse document frequency (*idf*).

The term frequency (*tf*) in the given document is simply the number of times a given term appears in that document. It is defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (4.1)$$

Where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j , and the denominator is the sum of number of occurrences of all terms in document d_j .

The inverse document frequency (*idf*) is regarded as a measure of importance of the term in the collection. It is defined as:

$$idf_i = \log \frac{N}{n_i} \quad (4.2)$$

⁸ Stop word list 1 which contains 429 words (available from: <http://truereader.com/manuals/onix/stopwords1.html> last accessed 19/3/2010)

Where N is the total number of documents in the collection, and n_i is the number of documents in which the term t_i appears.

Then $tf.idf$ is defined as:

$$(tf \cdot idf)_{i,j} = tf_{i,j} \cdot idf_i \quad (4.3)$$

5.2.2 Named Entity Recognition (NER)

Previously I have discussed the reviewed works of NER in TDT and *i*TDT in Chapter 2. It is important to understand how the NER system works. ANNIE is an information extraction component of GATE (General Architecture for Text Engineering) which I use for its accurate entity, pronoun and nominal co-references extraction (Cunningham et al., 2002). ANNIE is also chosen as an example of a typical NER system because it is freely available to the research community. In addition, the named entity types are a subset of the Message Understanding Conferences (MUC) types (Cunningham et al., 2002). ANNIE recognises the standard MUC entity types of Person, Location, Organisation, Date, Time, Money and Percent, plus the additional entity types Address and Identifier. ANNIE is able to recognise proper nouns, person, organisations, dates and locations. Based on ANNIE's capability, therefore I am not building a NER system and instead using the existing system to recognise named entities in a document. NER in ANNIE is based on gazetteer⁹ lists and JAPE (Java Annotation Patterns Engine) rules such as those depicted in Figure 5.2.

⁹ The GATE *ListGazetteer* are plain text files, with one entry per line. Each list represents a set of names, such as names of countries, cities, organisations, days of the week and others. Below is a small section of the list for *Country*:

...
Afghanistan
Afrique
Albania
...

```

Rule: Company1
Priority: 25
(
  ( {Token.orthography == upperInitial} )+ //from tokeniser
    {Lookup.kind == companyDesignator} //from gazetteer lists
  ) :match
-->
  :match.NamedEntity = { kind=company, rule="Company1" }

```

Figure 5.2: Rules in ANNIE for *Company*

Based on the *Company1* rule, ANNIE will identify a token start with a capital letter and will search for the company designator that matches from the gazetteer list. This rule will capture term such as *Acer*, *IBM*, *AOL* or *P&G* as a *Company*. If the token matches with the rule, ANNIE will recognise it under Organisation.

5.2.3 Document Clustering

The documents are clustered using Single Pass Clustering, a technique that has been proven to be reasonably effective for TDT (Papka and Allan, 1998; Eichmann & Srinivasan, 2002). Single-pass clustering, as the name suggests, requires a single, sequential pass over the set of documents it attempts to cluster. The algorithm is shown in Figure 5.3.

```

for each document d in the sequence loop
  1. find a cluster c that maximises  $\cos(c, d)$ ;
  2. if  $\cos(c, d) > t$  then include d in c;
  3. else create a new cluster whose only document is d;
end loop.
t is the similarity threshold value, which is usually derived experimentally.

```

Figure 5.3: Single-pass algorithm

The algorithm classifies the next document in the sequence according to a condition on the similarity of function employed. At every stage, the algorithm decides on whether a newly seen document should become a member of an already defined cluster or the centre of a new one. In its most simple form, the similarity function gets defined on the basis of just some similarity measure between document-feature vectors. In this work the similarity between two centroid vectors and between a document and a centroid vector are computed using the cosine measure:

$$similarity = \cos(\theta) = \frac{c \cdot d}{\|c\| \|d\|}. \quad (4.4)$$

5.3 *i*Event Interface

The reviewed work of the *i*TDT interface discussed in Chapter 2 has affected the design of *i*Event. *i*Event is composed of three components which are Cluster View (CV), Document View (DV) and Term View (TV). In this section I describe the design of *i*Event and discuss its components and features.

The layout and the order of the component displayed on the interface begins from the Cluster View followed by the Document View and finally the Term View. Cluster View is displayed on top of the interface as the main component since this is the starting point where users are presented with a large amount of information for rapid interpretation. Visualising the cluster based on the size and the density of the documents might help them to identify the important and related cluster based on the task given. Cluster View allows the users to browse the whole collection before they narrow their search to a specific cluster. That is the reason why the order of the Document View is after the Cluster View. The Document View allows the users to view the whole document in a cluster with the specific timeline. I believe it provides an effective form of presentation and a very fast graphical overview of the information that a cluster contains. Document View generates an interactive timeline displaying the major events and uses it as a

browsing interface to a document collection contained in a cluster. Finally the Term View is displayed at the bottom of the interface to be more specific on the terms contained in the cluster. Users get the whole view of the corpus before they receive specific information of the documents and the terms occurring in a cluster. The sequence or the ordering of the components on *iEvent* helps users to narrow down their browsing and to be focused in their searching. Thus it helps them to perform the TDT tasks.

iEvent has two settings. Setup 1 (Figure 5.4) is the baseline setup that uses keywords and Setup 2 (Figure 5.5) is the experimental setup that uses NER.

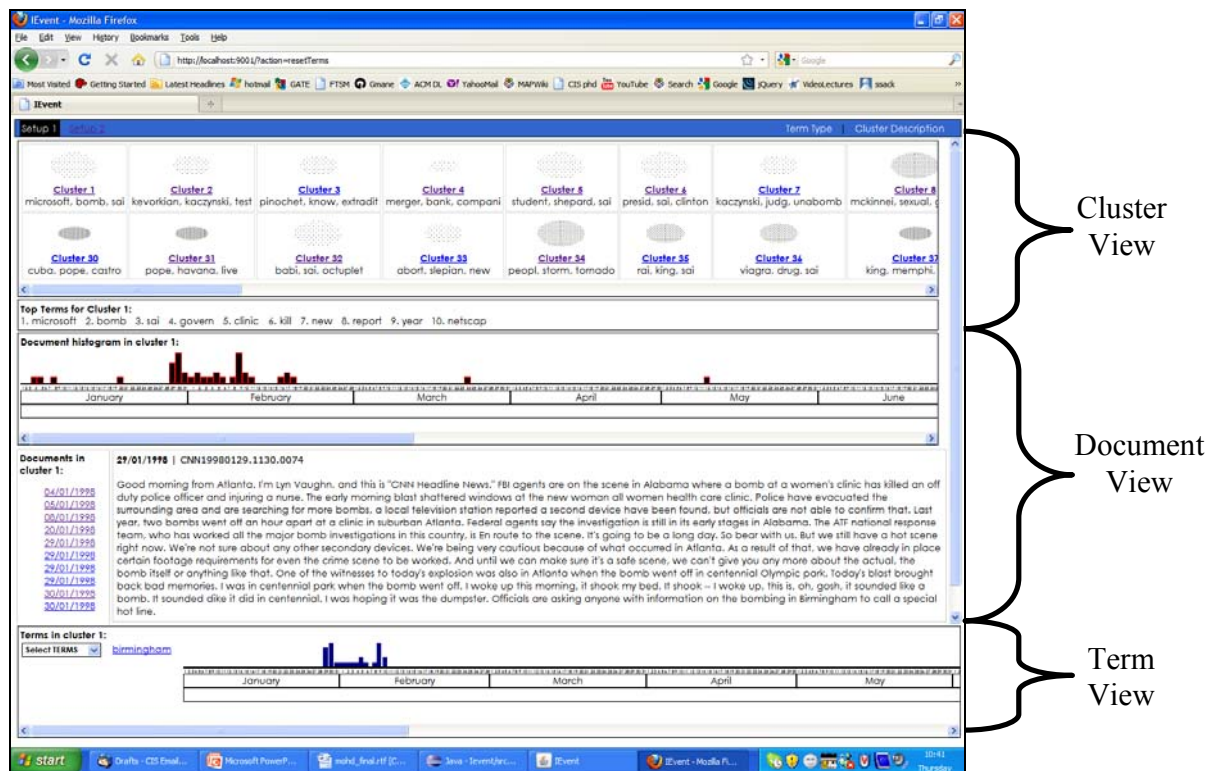


Figure 5.4: Keywords Setup (Setup 1)

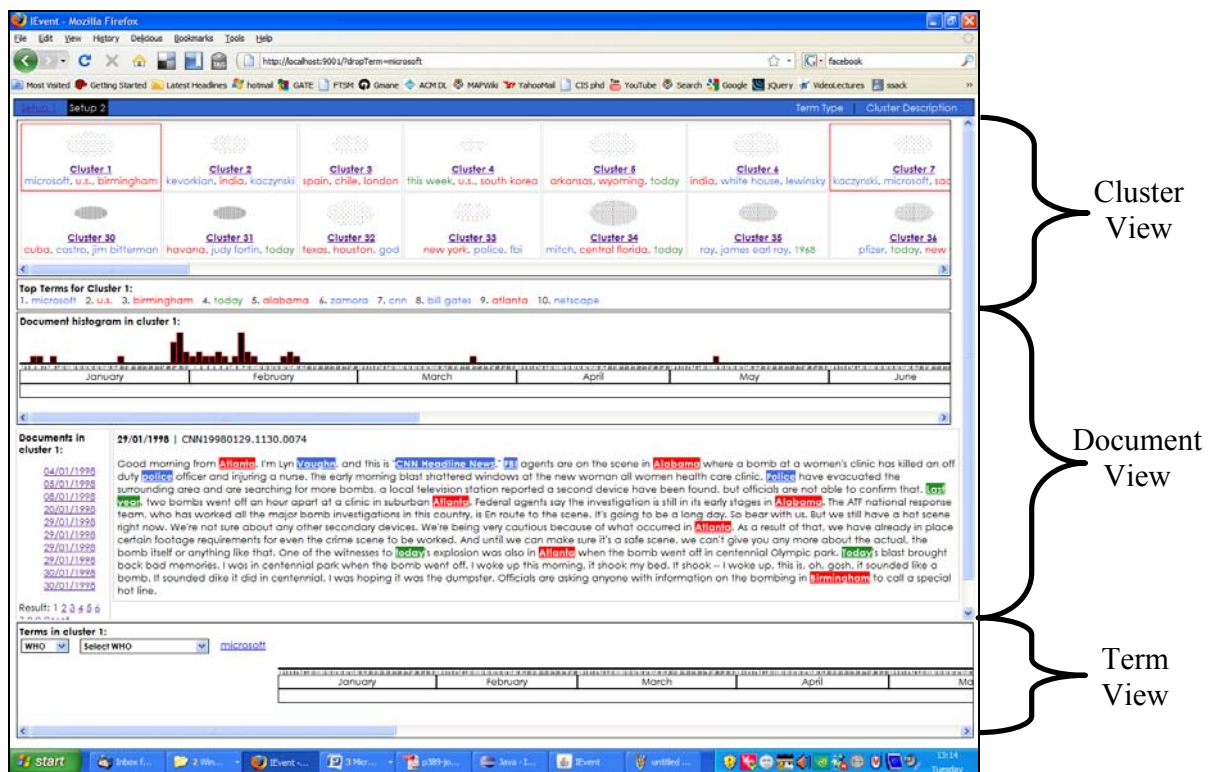


Figure 5.5: Named entities Setup (Setup 2)

Users are given the same amount of data i.e. the document and the clusters on the interface. Setup 1 (baseline setup) uses the same interface components with the features as the Setup 2 (experimental setup), but differs in using keywords instead of named entities. Therefore the user can switch whether they would like the named entities to be highlighted on the interface using Setup 2. I used Cascading Style Sheets (CSS) to differentiate three types of named entities with different colours assigned as shown in Table 5.1.

Type of named entities	Example
WHO: person, organisation	Anna, IBM
WHERE: location	Glasgow, Roskilde
WHEN: date	October, tomorrow, 2009

Table 5.1: Type of named entities

5.3.1 Cluster View (CV)

The Cluster View displays information related to the size and the density of a cluster; and the ten most frequently named entities in a cluster. The clusters are visualised based on the size and the density as shown in Figure 5.6.

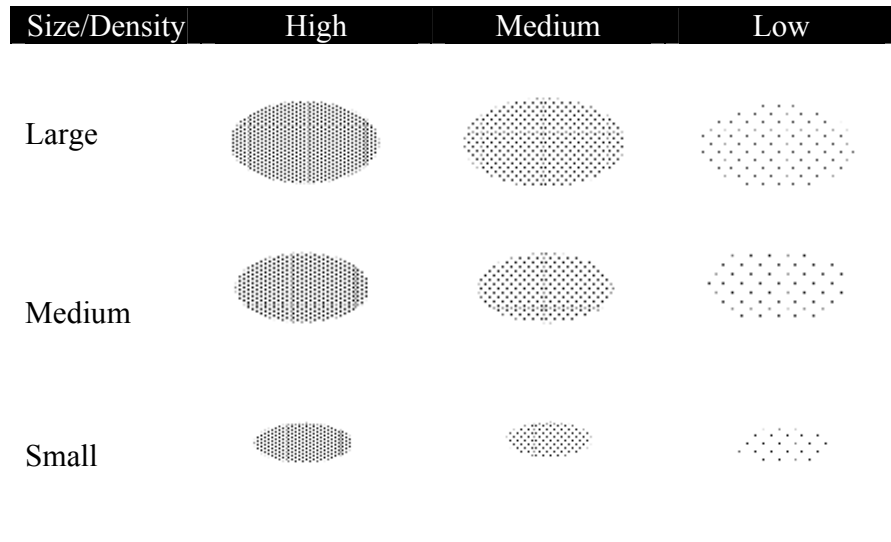


Figure 5.6: Size and density of the clusters

Clusters with a large size and high density contain lots of documents which have appeared over a short period of time, therefore they are supposed to represent very important events. On the other hand, clusters with a small size and low density contain a small number of documents which have appeared over a long period of time, thus presenting recurring but relatively unimportant events. Cluster visualisation is intended to help the user to make a rapid interpretation of a topic. It should be noted that given the difficulty in story segmentation, sometimes a cluster with a large size and low density might indicate the presence of more than one topic in the cluster.

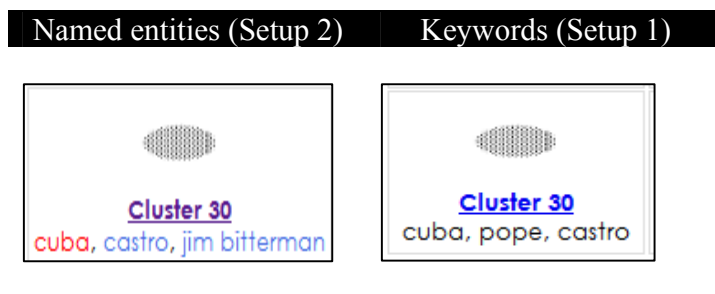


Figure 5.7: Cluster labelling of the two approaches

Clusters are labelled using the three most frequently named entities, as shown in Figure 5.7. For example, Cluster 30 is labelled using named entities in Setup 2 (experimental setup) instead of using keywords in Setup 1 (baseline setup).

When the user clicks on the cluster, additional information on the ten most frequently named entities in that cluster is presented, as shown in Figure 5.8. The difference of this feature between setups is that users are provided with named entities in Setup 2 (experimental setup) instead of using keywords in Setup 1 (baseline setup).

*i*TDT works such as Event Organizer (Allan et.al, 2005), TDTLighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000) have labelled the cluster or the topic using the ten most frequent terms. I decided to use this technique and to provide the ten most frequent terms in a separate box and labelled the cluster using the three most frequent terms due to the space issue. Single Pass Clustering used in this work as being reported in Section 5.2.3 (Document Clustering) has generated 57 clusters. The clusters are displayed based on the size and the density and labelling them with the three most frequent terms is appropriate to make sure Cluster View could handle the amount of information to be displayed.

Approach	Top ten terms
Named entities (Setup 2)	Top Terms for Cluster 30: 1. cuba 2. castro 3. jim bitterman 4. havana 5. varela 6. pope john paul ii 7. felix varela 8. havana university 9. 1959 10. friday
Keywords (Setup 1)	Top Terms for Cluster 30: 1. cuba 2. pope 3. castro 4. cuban 5. univers 6. havana 7. presid 8. meet 9. father 10. varela

Figure 5.8: Top ten terms of the two approaches

I believe these features are useful in TDT tasks since it provides information on the most frequent named entities or terms that occur in a specific cluster.

5.3.2 Document View (DV)

The Document View displays information about the document timeline and the documents contained in a cluster. The document timeline is displayed in a histogram form to show the occurrence and the document frequency for a specific date. The height of the histogram indicates the number of documents that occurred on that specific date in a cluster. This feature is an attempt to support the user in analysing the discourse or the information flow in a press article.

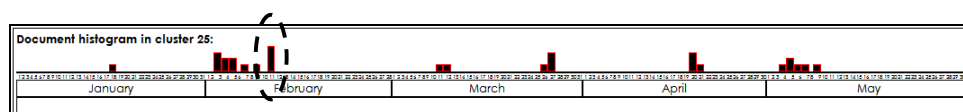


Figure 5.9: Document histogram for topic *Cable Car Crash* (Topic 20019)

Figure 5.9 is an example of a document timeline for topic Cable Car Crash (Topic 20019). It has a low density of documents indicating that the documents appear over a long period of time. The documents occurring in February report on the crash itself with

the highest documents occurring on February 11th. A month later the results of the investigation start to appear and the document in March refers to the investigation of the crash including the legal proceedings. Finally, documents in April and May, contain information related to the court case and the outcome of the crash.

Timelines are a useful way to present information that has a temporal dimension. Journalists often generate timelines to describe the course of events. This will be evaluated to prove that automatically generated timelines could prove invaluable for navigating the results of a TDT system and for interactive TDT. The timeline feature is offered both in the Document View and in the Term View. Users would be able to see the occurrence of the document and named entities within the timeline in a histogram form for each cluster.

Users would also be able to see the document content, with named entities highlighted, as shown in Figure 5.10.

Approach	Document content
Named entities (Setup 2)	<p>04/01/1998 CNN19980104.1130.0453</p> <p>and in Sacramento, California, opening statements are set for tomorrow in the trial of Theodore Kaczynski, four mail bombings blamed on the unabomber.</p>
Keywords (Setup 1)	<p>04/01/1998 CNN19980104.1130.0453</p> <p>and in Sacramento, California, opening statements are set for tomorrow in the trial of Theodore Kaczynski, four mail bombings blamed on the unabomber.</p>

Figure 5.10: Document content of the two approaches

The difference of this feature between setups is that the named entities in the document content are highlighted with different colours in Setup 2 (experimental setup), meanwhile Setup 1 (baseline setup) displays the document content without highlighting the named entities. I did not highlight the document content in Setup 1 because it will be a distraction in viewing the document since most of the terms have been highlighted.

5.3.3 Term View (TV)

The Term View displays information on the occurrence of the named entities within the timeline in a cluster. The timeline is displayed in the form of a histogram to show the named entities' occurrence and their frequency for a specific date. The histogram with the timeline shows the relevant score of named entities using term frequency (tf). Figure 5.11 shows an example of the highest score for named entity *Italy* occurring on February, 11th.

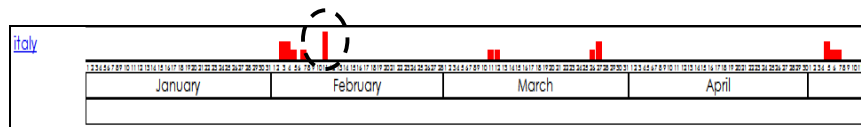


Figure 5.11: Histogram and the timeline for named entity *Italy* (Topic 20019)

The timeline feature provides journalists with the whole view of named entities occurrences in the cluster, as depicted in Figure 5.12. For example the latest occurrence of named entity *James Earl Ray* is on the 23rd and 24th April indicates his death based on the timeline. This is helpful in providing information about when the event occurred and supporting the new event detection task. This feature also helps the user in the Topic Detection task by presenting information about the latest occurrence of a named entity from the timeline.

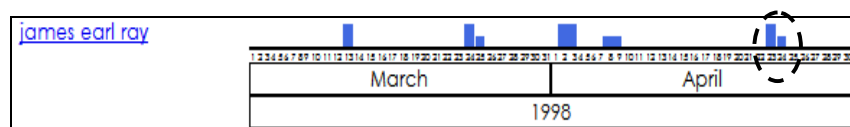


Figure 5.12: Histogram and the timeline for named entity *James Earl Ray* (Topic 20056)

5.4 Example of the use of *iEvent* in TDT tasks

In this section, I discuss the process of performing Topic Tracking and Topic Detection tasks using both setups.

5.4.1 Tracking Task

- a. Users receive the topic summary and they then have to track the related clusters.
- b. Then they browse the clusters displayed in the **Cluster View**
 - i. They use the features in Cluster View such as the cluster visualisation to get information on the size and the density of the documents contained in a cluster.
 - In Setup 2, clusters are labelled with the three most frequent named entities and additional information on the top ten named entities should provide them more information about a cluster.
 - In Setup 1, clusters are labelled with the three most frequent terms and additional information on the top ten terms.
 - ii. Once users have identified the related clusters, they will investigate by clicking on it.
- c. In **Document View**, the list of documents contained in the selected cluster is sorted by date and the document id is labelled with the format such as 04/01/1998 (dd/mm/yy).
 - i. Users can use the information such as the document histogram with the timeline to receive information on the occurrence and the frequency of the document for a specific date. The document id will be highlighted if they click on the document histogram.
 - ii. Users also receive a snippet showing the first 100 characters of the document content when they hover the mouse over the document id. The dataset used in the experiment as reported in Section 3.4.1 (Experimental Data) contained

documents with short stories, therefore displaying the first 100 characters to be sufficient.

- iii. Once the related documents have been identified, they clicked on each one to receive the document's content and started performing the Reporting task.
 - In Setup 2, the named entities are highlighted in the document content with different colours.
 - In Setup 1, users received the document content without any term being highlighted.
- d. Finally in **Term View**, Setup 2 will list out named entities that have been categorised into three groups; *Who*, *Where* and *When* type of named entity. Meanwhile in Setup 1, users received a list of terms.
 - In Setup 2, users could use the information such as the named entity histogram and the timeline to receive information on the number of the named entity occurring on a specific date. The document id will be highlighted if they click on the named entity histogram.
 - In Setup 1, users can use the information such as the term histogram and the timeline to receive information on the number of the terms occurring on a specific date. The document id will be highlighted if they click on the term histogram.

5.4.2 Detection Task

- a. Users receive a specific cluster from which they have to identify the topic being discussed in the cluster.
- b. They go to the specific cluster displayed in the **Cluster View**.
 - i. They use the features in Cluster View such as the cluster visualisation to get information on the size and the density of the documents contained in a cluster such as whether the cluster contains more than one topic.

- In Setup 2, they receive information about the three most frequent named entities and additional information on the top ten named entities. These should give them a hint about the topic being discussed in a cluster.
 - In Setup 1, they receive information about the three most frequent terms and additional information on the top ten terms.
- c. In **Document View**, they start to click and explore the documents contained in a cluster.
- i. Users can use the information such as the document histogram with the timeline to receive information on the frequency and the occurrence of the document on a specific date. The document id will be highlighted if they click on the document histogram.
 - ii. Users also receive a snippet showing the first 100 characters of the document content when they hover the mouse over the document id.
 - iii. Users click on the document id to receive the document content and detect the related topic. This is where they refer to the list of topics given during the Detection task to match with the information received from the interface.
 - In Setup 2, the named entities are highlighted in the document content with different colours.
 - In Setup 1, users receive the document content without any terms being highlighted.
- d. Finally in **Term View**, Setup 2 will list out named entities that have been categorised into three groups; the *Who*, *Where* and *When* types of named entity. Meanwhile in Setup 1, users receive a list of terms. The named entities/terms are sorted by the frequency. Users might use this information to identify the most frequent named entities/terms in that cluster to help them identify the topic.
- In Setup 2, users can use the information such as the named entity histogram and the timeline to receive information on the number of the named entity occurring for a specific date. The document id will be highlighted if they click on the named entity histogram.

- In Setup 1, users can use the information such as the term histogram and the timeline to receive information on the number of the term occurring for a specific date. The document id will be highlighted if they click on the term histogram.

5.5 Chapter Summary

iEvent incorporates some of the successful features from TDT interfaces into a single interface. It is designed to allow journalists to perform TDT tasks interactively. Each component of it is in line with the journalist tasks and relates to the TDT tasks. Cluster View helps the user to make a quick interpretation of a topic and to understand news in a relatively fast and efficient manner. Document View is an attempt to support the user in analysing the discourse or the information flow in a press article. I believe it provides an effective form of presentation and a very fast graphical overview of the information that a corpus contains. Finally, the Term View is helpful in providing information related to when the event occurred and in supporting the new event detection task. It also provides the latest occurrence of a named entity from the timeline.

I believe that the components and specific features of *iEvent* will contribute to assisting journalists in performing the TDT tasks, thus it is important to test whether *iEvent* can, in reality, enable journalists to perform well. In the next chapter I provide an evaluation of *iEvent* and investigate which components or features of *iEvent* are effective and in which TDT tasks.

Chapter 6

Evaluation of *iEvent*

6.1 Introduction

The literature reviewed in Chapter 2 revealed that none of the interactive Topic Detection and Tracking (*iTDT*) interfaces has been properly evaluated. Thus the *iEvent* (Interactive Event Tracking System) evaluation is important in identifying the good components and features of an *iTDT* interface design. The evaluation also aims to prove whether the implemented features of *iEvent* interface are effective in facilitating the performance of TDT tasks by professionals. This has motivated me to conduct a user experiment and to evaluate the *iEvent* interface with journalists performing TDT tasks. In this chapter I describe a pilot test, the experimental methodology of the user experiment and finally present the results.

6.2 Organization of *iEvent* Evaluation

The structure of *iEvent* evaluation starts with the discussion on the General Findings (Section 6.3) of *iEvent* such as the participants' likeability of *iEvent*; their topic familiarity and topic interest; and the comparison of participant's opinions between news network tools that participants used (i.e. Google, BBC, CNN) and *iEvent*.

Next the participants' performance in the Tracking task (Section 6.4) is analysed in terms such as the amount of successful Tracking tasks and participants' opinion of *iEvent* interface. I go on to analyse their performance in the Reporting task (sub activities in the Tracking task) such as how much they wrote. I then investigate the particular features of *iEvent* that participants perceived as useful, effective, helpful and interesting.

Then I examine the participants' performance in the Detection task (Section 6.5) such as the amount of successful Detection tasks, the participants' opinion on the easiness of detecting the topics and the usefulness of features used to perform the task.

Finally I compare and explain which features of *iEvent* will better facilitate the participants in performing both tasks.

6.3 General Findings

This section presents the general findings of *iEvent*. Participant performance was analysed to identify the effectiveness of *iEvent* interface in facilitating them to perform the Tracking and the Detection tasks. During the experiment 240 tasks were performed. 160 (66.67%) of these tasks were Tracking while the remaining (33.33%) of tasks were Detection.

Findings revealed that 70% of the participants liked *iEvent* and 50% of the participants prefer to use *iEvent* in both tasks. A possible explanation for these results might be the participants' success in performing both tasks (see Figure 6.2, Figure 6.10). 20% of participants disliked *iEvent* and 10% of participants were not sure. Those who disliked *iEvent* were all journalists that had an average age of 30-40 years and average working experience of more than 10 years. From the interview session, these participants had previously used news network tools such as PressDisplay.com, PaidContent.org and Google Fast Flip. Thus they had a high expectation when using *iEvent*. 10% were not sure, although they mentioned some interesting features of *iEvent*. However they disliked the fact that they had to scroll and mouse over the Cluster View to find the topic in the Tracking task.

Participants were asked about their topic familiarity and topic interest before they started using *iEvent*. Findings revealed that there was no statistical significance difference between topics and topic familiarity (Mann-Whitney Test, $p=0.483$). The participants were not familiar with the topics given in the Tracking task (mean=2.01 sd=1.03). There were also no statistically significant difference between the participants and their topic interest (Mann-Whitney Test, $p=0.842$). Their topic interest was average (mean=3.27 sd=1.09). This is a good indication of the experiment since the participants are not affected by external factors such as their topic familiarity and topic interest.

A Wilcoxon Signed Ranks test proved that there was a statistical significance difference in both topic familiarity and topic interest before and after using *iEvent* as shown in Table 6.1. The mean for topic familiarity and topic interest was increased after using *iEvent*.

	p-value	Mean	
		Before	After
Topic familiarity	0.000	2.01 (sd=1.025)	3.26 (sd=1.012)
Topic interest	0.000	3.27 (sd=1.092)	3.63 (sd=0.976)

Table 6.1: The mean for topic familiarity and topic interest

Table 6.2 shows that there was an increasing percentage (5 times higher) for participants who were familiar with the topic before (8%), and after (46%) using *iEvent*. The percentage decreased for participants who were not familiar with the topic before (69%) and after (27%) using *iEvent*. 69% of participants were not familiar with the topic because the collection used was in year 1998 (TDT2 and TDT3 corpus). Thus it supports the evaluation that *iEvent* influenced their topic familiarity and topic interest. If participants were given more recent topics, they might have been familiar with it and probably possess better knowledge of the topics that would have influenced their performance in the Tracking task.

		Scale (%)					(%)	
		1	2	3	4	5	(-)ive	(+)ive
Topic Familiarity	Before	40.0	29.4	22.5	6.3	1.9	69	8
	After	2.5	24.4	26.9	36.9	9.4	27	46
Topic Interest	Before	8.8	14.4	26.9	41.3	8.8	23	50
	After	2.5	10.0	28.1	41.3	18.1	13	59

(-)ive=scale 1, 2; (+)ive=scale 4,5 (scale from 1 to 5, higher=better).

Table 6.2: Topic familiarity and topic interest

For the topic interest, there was an increasing percentage for participants who were interested in the topic before (50%) and after (59%) using *iEvent*. Meanwhile the percentage of participants who were not interested in the topic before decreased to 10% after using *iEvent*.

A Mann-Whitney Test confirmed that there was no statistical significance difference ($p=0.492$) in topic interest before using *iEvent* across setups. However there was a statistical significance difference in topic interest after using *iEvent* across setups (Mann-Whitney Test, $p=0.003$). The participants were more interested with a topic in the Tracking task after using Setup 2 (mean=3.81 sd=1.032). Participants found using Setup 2 had enhanced their topic interest. It is apparent from Table 6.3 that there is a ratio of 7 participants to 1 who found that they were more interested in a topic after using Setup 2. They found that using Setup 2 of *iEvent* had significantly enhanced their topic interest with 46.3% of participants agreeing that they were interested (scale 4) with a topic.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	0.0	15.0	37.5	36.3	11.3	15.0	47.5	3:1
Setup 2	5.0	5.0	18.8	46.3	25.0	10.0	71.3	7:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.3: Topic interest (after) across setups

These results indicate that the participants were more familiar with the topics in the Tracking task after using *iEvent*. They were also more interested in the topics in the Tracking task after using Setup 2 of *iEvent*.

The participants were given an entry questionnaire before they performed the Tracking and Detection task. They were asked to list out the news network, tools or search engines used. Figure 6.1 shows that the participants mostly used Google (95%) and BBC news (90%) as their main news networks tools.

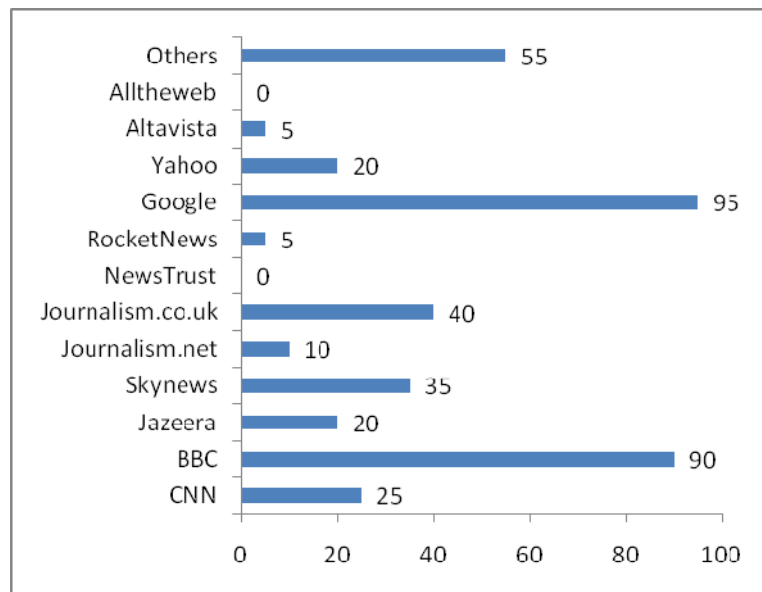


Figure 6.1: Percentage of the news networks tools used

Participants were also asked to rate their experience in using the news network tools as shown in Table 6.4.

Opinion	Mean (sd)	Scale (%)					(%)	
		1	2	3	4	5	(-)ive	(+)ive
Easy	4.05 (sd=0.759)	0.0	0.0	25.0	45.0	30.0	0.0	75.0
Relaxing	3.45 (sd=0.887)	0.0	0.0	10.0	50.0	25.0	0.0	75.0
Simple	3.10 (sd=0.912)	0.0	30.0	35.0	30.0	5.0	30.0	35.0
Satisfying	3.00 (sd=0.918)	0.0	35.0	35.0	25.0	5.0	35.0	30.0
Interesting	3.35 (sd=0.671)	0.0	10.0	45.0	45.0	0.0	10.0	45.0

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.4: Percentage of the participant opinions of the news network tools used

45% of participants found the news network tools that they used were easy (scale 4) (mean=4.05 sd=0.759). 50% of the participants found that the news network tools were relaxing (scale 4) (mean=3.45 sd=0.887). 35% of participants agreed that the news network tools were neither simple nor complex (scale 3) (mean=3.10 sd=0.912). Based

on participants' satisfaction, 35% of them were dissatisfied (scale 2) and found that the news network tools were average (scale 3) (mean=3.00 sd=0.918). Finally based on participants' interest, 45% of them found that the news network tools were average interesting (scale 3) and interesting (scale 4) (mean=3.35 sd=0.617). Participants interviewed mentioned the Google style of searching contributed to the ease of use of the news network tools, thus making the search process more relaxed and interesting.

6.4 Tracking Task

Several analyses were performed on the captured data. The following sections present the findings. First, the overall participants' opinions of *iEvent* are examined. Next, I investigate the participants' performance using *iEvent* in the Reporting task i.e. the amount of news written. Then, I investigate the features of *iEvent* that participants perceived as useful, effective, helpful and interesting.

6.4.1 Overall Opinions

The *iEvent* interface that participants perceived as easy, relaxing, simple, satisfying and interesting during the Tracking task was analysed as shown in Table 6.5. I also investigated the particular setups of *iEvent* that they perceived as easy, relaxing, simple, satisfying and interesting.

Opinion	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Easy	3.8	14.4	23.1	38.8	20.0	18.2	58.8	3:1
Relaxing	0.0	13.8	35.0	38.8	12.5	13.8	51.3	4:1
Simple	3.8	21.3	26.9	37.5	10.6	25.1	48.1	2:1
Satisfying	1.3	8.8	38.8	40.0	11.3	10.1	51.3	5:1
Interesting	0.0	6.9	31.9	26.9	34.4	6.9	61.3	9:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.5: Percentage of the participant opinions of *iEvent*

a. Easy

As shown in Table 6.4, a ratio of 3 participants to 1 found that *iEvent* was easy to use (mean=3.57 sd=1.079) with 38.8% of participants agreeing that it was easy (scale 4). During the interview session, the participants informed me that *iEvent* was easy because it has structured and clear components; Cluster, Document and Term Views, thus making it easy to use.

There was a statistical significance difference in participants' opinions (easy) across setups (Mann-Whitney Test, p=0.004). 45.0% of participants agreed that Setup 2 (mean=3.85 sd=0.828) was easy (scale 4). Interestingly there were 67.5% of participants who found that Setup 2 was easier compared to 5% who found it difficult. This indicates that 14 participants found that using Setup 2 of *iEvent* made the Tracking task easier as shown in Table 6.6.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	7.5	23.8	18.8	32.5	17.5	31.3	50.0	2:1
Setup 2	0.0	5.0	27.5	45.0	22.5	5.0	67.5	14:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.6: Percentage of the participant opinion (easy) across setups

b. Relaxing

As shown in Table 6.4, 4 participants to 1 found that *iEvent* was relaxing (mean=3.50 sd=0.883). 38.8% of participants agreed that it was relaxing (scale 4). Participants interviewed once more associated the relaxing factor with the structured and clear components of *iEvent*, which also supported the perceived ease of using *iEvent* to perform the Tracking task.

There was a statistical significance difference in participants’ opinions (relaxing) across setups (Mann-Whitney Test, p=0.003). 41.3% of participants agreed that Setup 2 (mean=3.71 sd=0.860) was relaxing (scale 4). 60% of participants who found that Setup 2 was more relaxing compared to 7.5% who found it stressful. This indicates that 8 participants found using Setup 2 of *iEvent* makes the Tracking task more relaxing as shown in Table 6.7.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	0.0	23.8	18.8	32.5	17.5	23.8	50.0	2:1
Setup 2	0.0	7.5	32.5	41.3	18.8	7.5	60.0	8:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.7: Percentage of the participant opinion (relaxing) across setups

c. Simple

There were 2 participants to 1 who found that *iEvent* was simple (mean=3.30 sd=1.039) with 37.5% of participants indicating that it was simple (scale 4) as depicted in Table 6.4. Mann-Whitney Test confirmed that there was no statistical significance difference on participants’ opinion in simple (p=0.840) in conjunction with the setups.

The participants interviewed relate this opinion with the clear and structured components of *iEvent* but there were also suggestions to revise the layout of *iEvent* especially that

the Cluster View should be vertical instead of horizontal. Thus it seems that the layout issue would be interesting one for future work involving *iEvent*.

d. Satisfying

As shown in Table 6.4, there were 5 participants to 1 who found *iEvent* to be satisfying (mean=3.50 sd=0.854) with 40% of participants agreeing that it was satisfying (scale 4). A Mann-Whitney Test confirmed that there was no statistical significance difference on participants' opinion in satisfying ($p=0.500$) in conjunction with the setups.

I measure participants' satisfaction by analysing their agreement on enough information gathered during the Tracking task and the Reporting task results. I believe participants were satisfied with *iEvent* if they managed to perform the Tracking task by receiving enough information for a topic and they managed to report the story assigned by tracking the correct cluster. Participants were deemed to be satisfied if they found the information that they needed. Further analysis showed that 39.4% of participants agreed that they had gathered enough information using *iEvent* (mean=3.50 sd=1.082) during the Tracking task. 3 participants to 1 agreed that they had gathered enough information using *iEvent*. A Mann-Whitney Test also confirmed that there was no statistical significance difference in their agreement on enough information gathered ($p=0.113$) in conjunction with the setups. This indicates that the participants had gathered enough information using both setups.

Interestingly, the satisfaction factor was also related to the high percentage of correct clusters to be tracked. This provides strong evidence that *iEvent* mostly helped to facilitate the participants in tracking the correct cluster (mean=3.87 sd=0.49). Mann-Whitney Test also confirmed that there was no statistical significance difference in the number of correct clusters to be tracked ($p=0.160$) in conjunction with the setups. This indicates that the participants managed to track the correct clusters using both setups.

I classify the correctness of cluster as being tracked into four categories:

- i. none- where participants did not provide any information or they did not complete the task
- ii. wrong- where participants tracked the wrong cluster.
- iii. partially correct- where participants list out the minor cluster as their main finding.
- iv. correct- where participants list out the major cluster as their main finding.

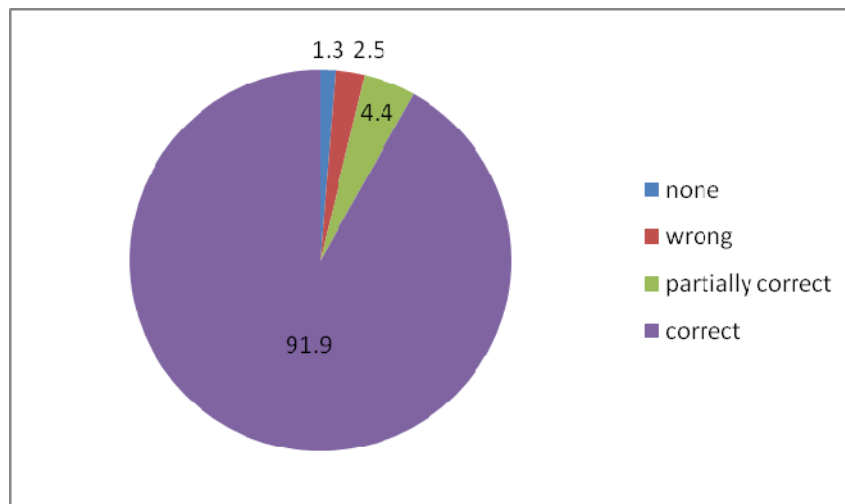


Figure 6.2: Percentage of successful Tracking task

The entire Tracking task was successful with 91.9% of tasks being correct and 4.4% being partially correct as shown in Figure 6.2. There were 2 participants (1.3%) who did not complete the task on topic National Tobacco Settlement. Participants were using Setup 2 (experimental setup) which displays information on named entities (e.g. Congress, Clinton) however they were looking for the term *tobacco*. This is the reason why participants spent the full 15 minutes allocated and were still not able to find the correct cluster. There were 4 participants (2.5%) who were wrong about the topic Mobil-Exxon Merger. Participants were confused with this topic when they were using Setup 1 (baseline setup) from the term *merge* which also highlights the cluster on topic Microsoft Merger. These uncompleted and wrong tasks represented just a small

percentage compared to the successful tasks. This proved that *iEvent* managed to facilitate the participants in performing well in the Tracking task.

e. Interesting

This opinion of *iEvent* received the highest ratio with 9 participants to 1 finding that *iEvent* was interesting (mean=3.89 sd=0.956). 34.4% of participant agreed that it was very interesting (scale 5) as depicted in Table 6.4.

There was a statistical significance difference in participants’ opinion (interesting) across setups (Mann-Whitney Test, $p < 0.05$). Setup 2 (mean=4.23 sd=0.779) received the highest percentage with 43.8% of participants agreeing that it was very interesting (scale 5). Surprisingly none of the participants found that Setup 2 was boring and 78.8% of participants found that it was more interesting as shown in Table 6.8. This indicates that the participants found using Setup 2 of *iEvent* makes the Tracking task more interesting than Setup 1.

	Scale (%)					(%)	
	1	2	3	4	5	(-)ive	(+)ive
Setup 1	0.0	12.5	43.8	18.8	25.0	12.5	43.8
Setup 2	0.0	0.0	21.3	35.0	43.8	0.0	78.8

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.8: Percentage of the participant opinion (interesting) across setups

One of the participants quoted that “This is a new paradigm of monitoring news in journalism and absolutely interesting”.

6.4.2 Reporting Task

This section reports the findings of participants' performance during the Reporting task as one of the sub activities of the Tracking task. I analyse the number of lines that participants wrote. This is an important measure to ascertain the effectiveness of *iEvent* in providing information to the participants. The more they wrote indicated that the participant received enough information and was able to deliver it in a written form. I also analysed the number of lines that participants wrote across setups. There was no statistical significance difference on the amount of news written in conjunction with the setups (Mann-Whitney Test, $p=0.434$) and no statistical significance difference on the amount of news written for different topics (Mann-Whitney Test, $p=0.202$). These indicate that the participants managed to write the amount of news equally using both setups and they managed to write the amount of news equally for every topic given in this experiment.

Findings revealed that the participants wrote on average nine lines using *iEvent* (mean=9.44 sd=6.455). There was a statistical significance difference in the amount of news written in conjunction with the type of participants (Mann-Whitney Test, $p<0.05$). The journalists (mean=7.09 sd=5.45) wrote less than the students (mean=11.79 sd=6.56) as shown in Figure 6.3. The reason for this was that the journalists were more selective and critical when writing news.

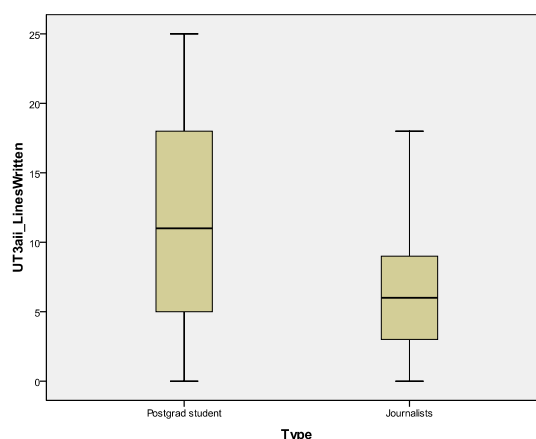


Figure 6.3: Boxplot of *how much* was written (scale 1-5) by each type of participant in the Tracking task

Table 6.9 shows an example of news written for the task on topic Mobil-Exxon Merger where they were asked to write an article on the outcome of the merger.

Student
<ul style="list-style-type: none"> • \$80 billion merger that would create the world's largest oil • Exxon and Mobil currently employ nearly 123,000 people. • About 12,000 people are expected to lose their jobs because of the merger. • Elimination of overlapping positions such as geologists, geophysicists, engineers plan to save \$2.8 billion a year. • New company may have to sell off some gas stations and refineries to satisfy government regulators • Combined company will most likely be forced to sell some of its holdings in order to appease antitrust regulators.
Journalist
<ul style="list-style-type: none"> • Economy: Jobless rate increased since 12,000 of Exxon Mobil employees were affected from the merger • Consumer: affordable prices since the merger provides more opportunities on energy exploration and new discoveries of oil and gas • Law: violating anti-trust law, merger creates monopoly and unfair business practices

Table 6.9: Comparison of news written on topic Mobil-Exxon Merger between the student and the journalist

As shown in Figure 6.4, *iEvent* also facilitated the participants to report the correct news (mean=3.80 sd=0.708). 91.3% of participants managed to report the correct news and interestingly none of the participants provided the wrong information.

There was no statistical significance difference on the amount of correct news written in conjunction with the setups (Mann-Whitney Test, $p=0.651$) and this indicated that the participants managed to write the amount of correct news equally using both setups. I classify the correctness of news written into four categories:

- i. none- where participants did not provide any information or they did not complete the task.
- ii. wrong- the news written did not match the topic.
- iii. partially correct- part of the news written matched the topic.
- iv. correct- the news written matched the topic.

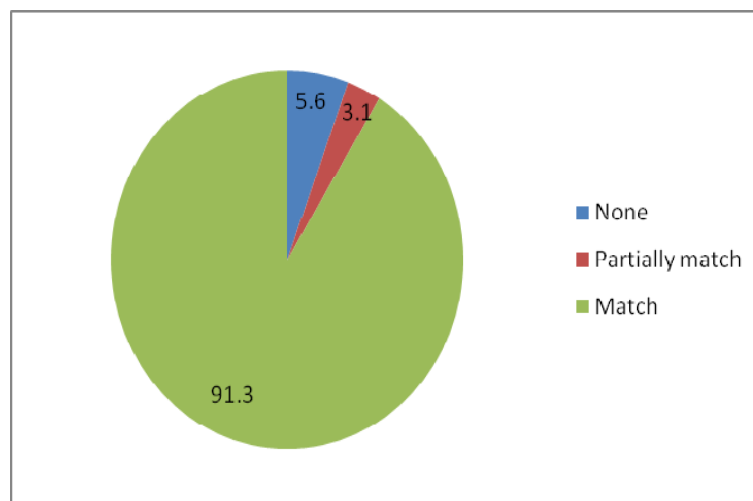


Figure 6.4: Percentage of correct news written

6.4.3 Features

I analysed each feature and assessed which setup of *iEvent* that participants perceived as useful, effective, helpful and interesting during the Tracking task.

6.4.3.1 Useful

In this section I analysed each feature of *iEvent* and assessed which setup participants perceived as useful during the Tracking task.

a. Useful features of *iEvent*

The highest ratio for this opinion was for the ‘CV: cluster visualisation’ feature. 10 participants to 1 found that this feature was useful in the Tracking task (mean=3.86 sd=1.008). 36.3% thought the ‘CV: cluster visualisation’ feature was useful (scale 4) as shown in Table 6.10. The size and the density of the clusters contained in this feature allow the participant to identify how many topics are in each cluster, such that clusters with large size and high density indicate a high number of documents where the distribution of the documents are over a long period of time.

There were also two features perceived as useful by the participants with ratio 6:1 which were the ‘DV: histogram with the timeline’ (mean=3.82 sd=1.192) and the ‘TV: histogram with the timeline’ (mean=3.79 sd=1.099). 44.4% thought the ‘TV: histogram with the time line’ feature was useful (scale 4) and 35% found the ‘DV: histogram with the timeline’ feature to be very useful (scale 5). These features allow the participant to see the document and the term occurrence for a specific date. A topic such as Jonesboro Shooting did mention the date 29th of April as the hearing case and using these features was an advantage in reporting the outcome of the trial.

FEATURES	Scale (%)					(-) ive	(+) ive	Ratio
	1	2	3	4	5			
CLUSTER VIEW (CV)								
cluster labelling	5.6	15.6	20.0	35.6	23.1	21.2	58.7	3:1
top terms	3.8	18.1	27.5	38.1	12.5	21.9	50.6	2:1
cluster visualisation	3.8	3.1	26.9	36.3	30.0	6.9	66.3	10:1
cluster description button	16.9	16.9	41.9	17.5	6.9	33.8	24.4	1:1
DOCUMENT VIEW (DV)								
histogram with the timeline	8.1	3.8	21.3	31.9	35.0	11.9	66.9	6:1
document content	1.3	11.9	23.8	38.8	24.4	13.2	63.2	5:1
TERM VIEW (TV)								
keyword approach	5.6	12.5	24.4	28.8	28.8	18.1	57.6	3:1
histogram with the timeline	6.3	6.3	16.3	44.4	26.9	12.6	71.3	6:1

(-) ive=scale 1, 2; (+) ive=scale 4,5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.10: Percentage of participants who perceived the features of *iEvent* as *useful* in the Tracking task

Moreover, 60% of the participants agreed that the ‘DV: document histogram with the timeline’ feature is a way to studying discourse analysis. Discourse analysis is important in journalism as it studies the information flow in a press article. These findings support the reason why participants gave a high score (scale 5) on the usefulness of this feature in Tracking task. In addition 40% of participants agreed that the document histogram with the timelines was the best feature of *iEvent*.

Further analyses of the interaction logs proved that the participants were not using the ‘CV: cluster description button’ feature, as there was a low activity (0.3%) in it. This is the reason why a high percentage (41.9%) of participants perceived it as being average (scale 3), as shown in Table 6.3.

There was a statistical significance difference on the ‘CV: cluster visualisation’ feature (Mann-Whitney Test, $p=0.002$) and the ‘DV: histogram with the timeline’ feature (Mann-Whitney Test, $p<0.05$) between students and journalists. These 2 features were

significantly more popular among students compared to journalists. Students found the ‘CV: cluster visualisation’ feature was useful (scale 4) and the ‘DV: histogram with the timeline’ was very useful (scale 5), as shown in Figure 6.5.

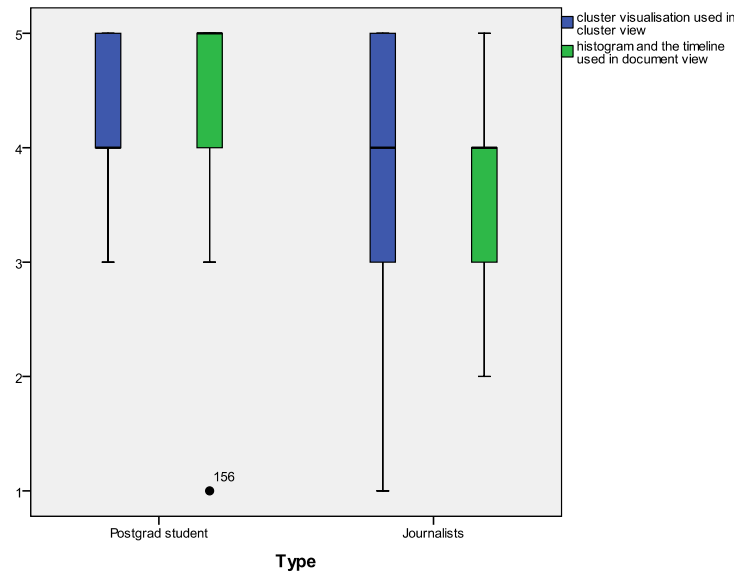


Figure 6.5: Boxplot on the *usefulness* (Scale 1-5) of the features for type of participants

b. Useful features of *iEvent* across setups

There was a statistical significance difference on the ‘CV: cluster labelling’ feature across setups (Mann-Whitney Test, $p < 0.05$). The ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* was more useful (mean=3.94 sd=0.919) compared to Setup 1 (mean=3.16 sd=1.267). 9 participants to 1 found that ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* was perceived as significantly useful with 38.8% of participants agreeing that it was useful (scale 4) as shown in Table 6.11.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	11.3	23.8	17.5	32.5	15.0	35.0	47.5	1:1
Setup 2	0.0	7.5	22.5	38.8	31.3	7.5	70.0	9:1

(-) ive= scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.11: ‘CV: cluster labelling’ feature across setups perceived as useful

This indicates that the participants found ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* more useful than Setup 1.

6.4.3.2 Effective

In this section I analysed each feature of *iEvent* and assessed which setup the participants perceived as effective during the Tracking task.

a. Effective features of *iEvent*

11 participants to 1 found that the ‘DV: document content’ feature was effective in the Tracking task (mean=4.41 sd=0.968). It can be seen from the data in Table 5.15 that the most striking results were that 45% found the ‘DV: document content’ feature very effective (scale 5). Further analyses on the interaction logs among the successful Tracking tasks proved that there was high activity using the ‘DV: document content’ feature with 71.4% of participants using it. This indicates that this feature was effective in facilitating the participant in tracking the correct cluster.

There were also two further features perceived as effective by the participants. 7 to 1 participants found that the ‘DV: histogram with the timeline’ (mean=3.97 sd=1.096) was effective. Moreover 39.4% found this feature to be very effective (scale 5). 6 participants to 1 found that the ‘CV: cluster visualisation’ feature (mean=3.73 sd=0.951) was effective too with 36.9% rating it as effective (scale 4).

FEATURES	Scale (%)							Ratio
	1	2	3	4	5	(-) ive	(+) ive	
CLUSTER VIEW (CV)								
cluster labelling	1.9	19.4	28.8	33.1	16.9	21.3	50	2:1
top terms	0.6	14.4	25.0	37.5	22.5	15	60	4:1
cluster visualisation	1.3	8.1	30.6	36.9	23.1	9.4	60	6:1
cluster description button	13.1	18.8	46.9	15.6	5.6	31.9	21.2	1:1
DOCUMENT VIEW (DV)								
histogram with the timeline	4.4	5.6	18.1	32.5	39.4	10	71.9	7:1
document content	1.3	5.6	16.3	31.9	45.0	6.9	76.9	11:1
TERM VIEW (TV)								
keyword approach	3.8	8.1	31.9	35.6	20.6	11.9	56.2	5:1
histogram with the timeline	6.9	13.8	13.8	41.3	24.4	20.7	65.7	3:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.12: Percentage of participants who perceived the features of *iEvent* as *effective* in the Tracking task

Again the ‘CV: cluster description button’ feature has a high percentage (46.9%) where it was perceived by the participants as average (scale 3) as shown in Table 6.12. This explains the fact that they did not use it as much as the ‘DV: document content’ feature.

There was a statistical significance difference in perception of the ‘CV: cluster visualisation’ feature (Mann-Whitney Test, $p=0.001$) and ‘DV: histogram with the timeline’ feature (Mann-Whitney Test, $p<0.05$) between students and journalists. The ‘CV: cluster visualisation’ feature was popular among the students as shown in Figure 6.6. The students interviewed mentioned that it was effective since it gave them quick information on the number of documents and the density. They mentioned that clusters with large size and high density had more than 1 topic so they preferred to investigate on the clusters with medium size with medium or high density.

The ‘DV: histogram with the timeline’ was popular among the journalists. The journalists interviewed claimed that the ‘DV: histogram with the timeline’ was effective since they were critical when looking for a very specific information. This feature allows them to answer the question on ‘*when was the event?*’

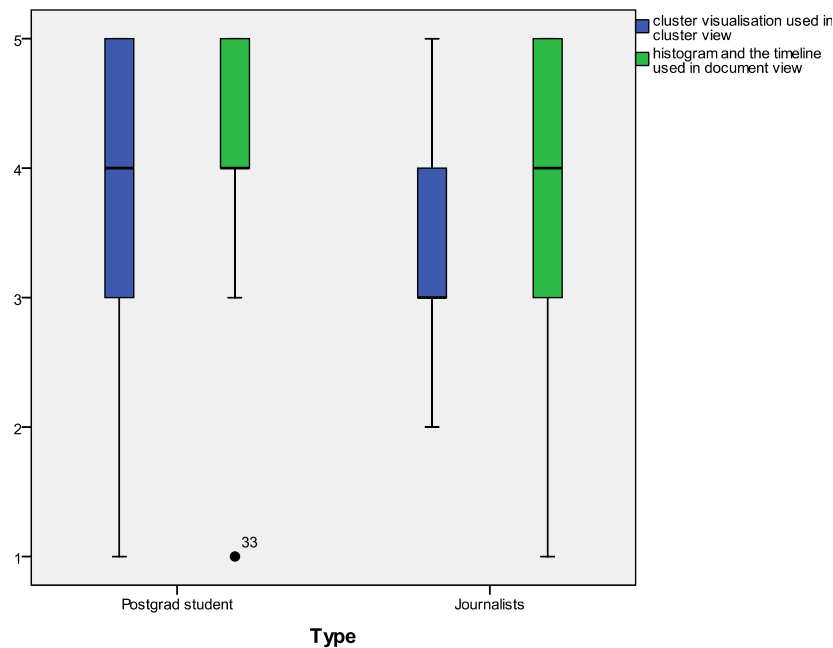


Figure 6.6: Boxplot on the *effectiveness* (Scale 1-5) of the features for type of participants

b. Effective features of *iEvent* across setups

There was a statistical significance difference on the ‘CV: cluster labelling’ feature across setups (Mann-Whitney Test, $p=0.008$). The ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* was more effective (mean=3.66 sd=0.927) compared to Setup 1 (mean=3.21 sd=1.110).

5 participants to 1 found that ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* was perceived as significantly effective with 41.3% of participants agreed that it was effective (scale 4) as shown in Table 6.13.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	3.8	26.3	30.0	25.0	15.0	30.0	40.0	1:1
Setup 2	0.0	12.5	27.5	41.3	18.8	12.5	60.0	5:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.13: ‘CV: cluster labelling’ feature across setups perceived as effective

This indicates that the participants found ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* more effective than Setup 1.

6.4.3.3 Helpful

In this section I analysed each feature of *iEvent* and assessed which setup participants perceived as helpful during the Tracking task.

a. Helpful features of *iEvent*

There were 12 participants to 1 who found that the ‘TV: histogram with the time line’ feature was helpful in the Tracking task. The participants interviewed mentioned that they could see the specific occurrence for a specific term. The topic *Jonesboro Shooting* for example, allowed them to scan the timeline for significant terms such as *Mitchell Johnson* and *Andrew Golden*. Thus 38.1% of the ‘TV: histogram with the time line’ feature was perceived to be very helpful (scale 5) as shown in Table 6.14. There were three features that were perceived to be helpful by 6 participants to 1. They were the ‘CV: top terms’ feature (mean=3.77 sd=0.992), the ‘CV: cluster visualisation’ feature (mean=3.71 sd=0.948) and the ‘DV: document content’ feature (mean=3.90 sd=1.083).

FEATURES	Scale (%)					(-) ive	(+) ive	Ratio
	1	2	3	4	5			
CLUSTER VIEW (CV)								
cluster labelling	1.3	15.0	27.5	41.3	15.0	16.3	56.3	3:1
top terms	0.6	9.4	31.3	30.0	28.8	10	58.8	6:1
cluster visualisation	0.6	9.4	31.3	35.6	23.1	10	58.7	6:1
cluster description button	10.6	19.4	45.6	16.9	7.5	30	24.4	1:1
DOCUMENT VIEW (DV)								
histogram with the timeline	1.3	11.3	26.9	30.6	30.0	12.6	60.6	5:1
document content	2.5	8.8	22.5	28.8	37.5	11.3	66.3	6:1
TERM VIEW (TV)								
keyword approach	0.6	15.0	20.6	38.1	25.6	15.6	63.7	4:1
histogram with the timeline	0.6	5.0	25.0	31.3	38.1	5.6	69.4	12:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.14: Percentage of participants who perceived the features of *iEvent* as *helpful* in the Tracking task

There was a statistical significance difference on the ‘TV: histogram with the time line’ feature (Mann-Whitney Test, $p=0.029$) between the topics. This feature was particularly popular for Topic 7 (German Train Derails), as shown in Figure 6.7, because it requires the participant to report the accident where timeline is an important feature to track the story of the accident, investigation and the consequences from the accident. Further analysis of the interaction logs for Topic 7 proved that the participants were using this feature more frequently for this topic, with 10.2% of activity compared to an average usage of 9%.

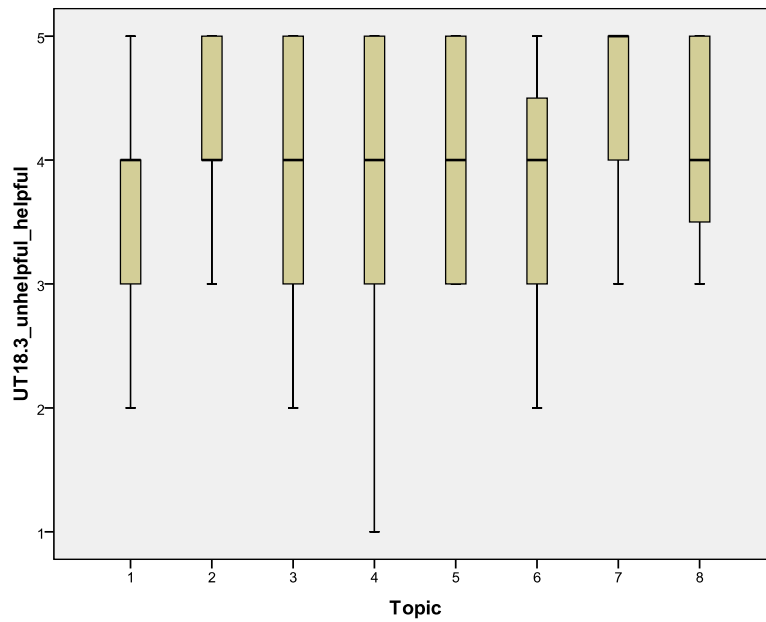


Figure 6.7: Boxplot on the *helpfulness* (Scale 1-5) of the ‘TV: histogram with the time line’ feature for Topics 1-7

There was a statistical significance difference on the ‘CV: cluster visualisation’ feature (Mann-Whitney Test, $p < 0.05$) and the ‘DV: histogram with the timeline’ feature (Mann-Whitney Test, $p < 0.05$) between students and journalists. The two features were more popular among the students compared to the journalists. They found the features were helpful (scale 4) as shown in Figure 6.8.

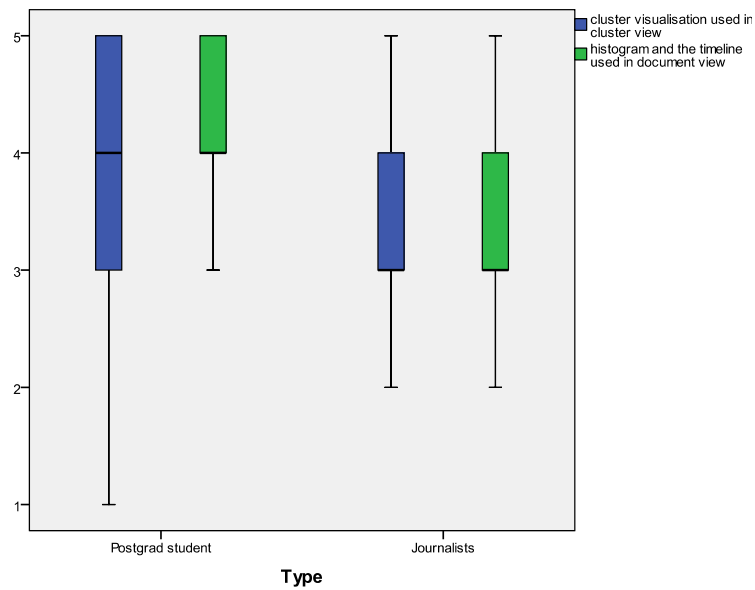


Figure 6.8: Boxplot on the *helpfulness* (Scale 1-5) of the features for type of participants

b. Helpful features of *iEvent* across setups

There was a statistical significance difference on two features across setups. The ‘CV: top terms’ feature (Mann-Whitney Test, $p=0.033$) and ‘TV: keyword approach’ features (Mann-Whitney Test, $p=0.011$) in Setup 2 of *iEvent* were more helpful than Setup 1.

17 participants to 1 found that the ‘CV: top terms’ feature in Setup 2 of *iEvent* was perceived as significantly helpful with 35.0% of participants agreeing that it was very helpful (scale 5) as shown in Table 6.15. 9 participants to 1 found that the ‘TV: keyword approach’ feature in Setup 2 was perceived as significantly helpful with 35.0% of participants agreeing that it was helpful (scale 4).

	Scale (%)					(-)ive (%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
CV: top terms								
Setup 1	1.3	15.0	30.0	31.3	22.5	16.3	53.8	3:1
Setup 2	0.0	3.8	32.5	28.8	35.0	3.8	63.8	17:1
TV: keyword approach								
Setup 1	1.3	22.5	17.5	41.3	17.5	23.8	58.8	2:1
Setup 2	0.0	7.5	23.8	35.0	33.8	7.5	68.8	9:1

(-) ive=scale 1, 2; (+) ive= scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.15: ‘CV: top terms’ and ‘TV: keyword approach’ features across setups perceived as helpful

This indicates that the participants found the ‘CV: top terms’ and ‘TV: keyword approach’ features in Setup 2 of *iEvent* more helpful than Setup 1.

6.4.3.4 Interesting

In this section I analysed each feature of *iEvent* and the particular setup that participants perceived as interesting during the Tracking task.

a. Interesting features of *iEvent*

It was apparent from Table 6.16, that there were three features that participants perceived as interesting which have a high ratio (more than 10:1) compared to other features. 14 participants to 1 found that the ‘DV: histogram with the timeline’ feature was interesting (mean=4.04 sd=0.983). The participants found that the ‘CV: cluster labelling’ feature (mean=4.04 sd=0.983) was interesting with 42.5% finding it very interesting (scale 5). During the informal interview session, the participants found this feature was very interesting because they received quick information on the topic using the 3 most frequent terms for the cluster.

11 to 1 participants found that the ‘CV: top terms’ feature (mean=4.03 sd=0.968) was interesting too, with 38.8% indicating they felt this feature was very interesting (scale 5).

FEATURES	Scale (%)					(-) ive	(+) ive	Ratio
	1	2	3	4	5			
CLUSTER VIEW (CV)								
cluster labelling	.6	5.6	25.6	25.6	42.5	6.2	68.1	11:1
top terms	1.3	5.6	20.6	33.8	38.8	6.9	72.6	11:1
cluster visualisation	4.4	6.9	36.3	30.0	22.5	11.3	52.5	5:1
cluster description button	8.1	11.3	56.9	13.1	10.6	19.4	23.7	1:1
DOCUMENT VIEW (DV)								
histogram with the timeline	.6	3.1	43.1	25.0	28.1	3.7	53.1	14:1
document content	4.4	8.8	27.5	26.9	32.5	13.2	59.4	5:1
TERM VIEW (TV)								
keyword approach	4.4	16.3	20.0	29.4	30.0	20.7	59.4	3:1
histogram with the timeline	1.3	12.5	16.3	38.8	31.3	13.8	70.1	5:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.16: Percentage of participants who perceived the features of *iEvent* as *interesting* in the Tracking task

There was a statistically significant difference on the ‘CV: cluster visualisation’ feature (Mann-Whitney Test, $p=0.049$) between students and journalists. This feature was popular among students since they not only found it effective (Figure 6.6) and helpful (Figure 6.8) but also interesting (scale 4) as shown in Figure 6.9.

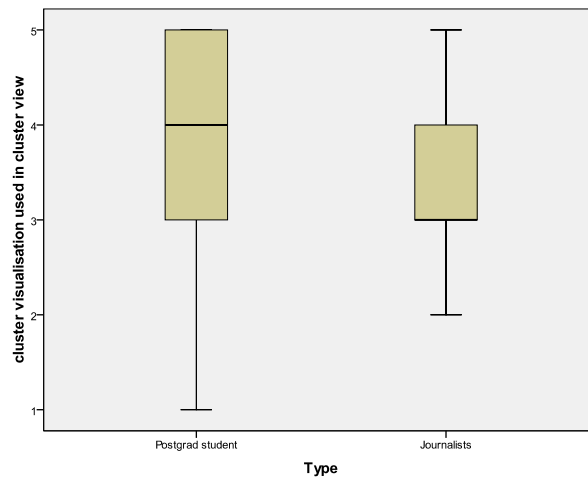


Figure 6.9: Boxplot on the *interestingness* (Scale 1-5) of the feature for type of participants

b. Interesting features of *iEvent* across setups

There was a statistically significant difference on four features across setups. The ‘CV: cluster labelling’ (Mann-Whitney Test, $p=0.033$), ‘CV: top terms’ (Mann-Whitney Test, $p=0.026$), ‘DV: document content’ (Mann-Whitney Test, $p=0.013$) and ‘TV: keyword approach’ features (Mann-Whitney Test, $p=0.035$) in Setup 2 of *iEvent* was more interesting than Setup 1.

Table 6.17 shows that there were two features in Setup 2 perceived as very interesting (scale 5) by the participants. They were the ‘CV: cluster labelling’ feature (48.8%) and the ‘CV: top terms’ feature (43.8%). There were also two further features in Setup 2 perceived as interesting (scale 4) by the participants; the ‘DV: document content’ feature (36.3%) and the ‘TV: keyword approach’ feature (37.5%). Surprisingly these four features received a high percentage (70%-82.5%) of participants who found it interesting.

	Scale (%)					(-)ive (%)		(+)ive (%)	
	1	2	3	4	5				
CV: cluster labelling									
Setup 1	1.3	8.8	36.3	17.5	36.3	10.0	53.8		
Setup 2	0.0	2.5	15.0	33.8	48.8	2.5	82.5		
CV: top terms									
Setup 1	2.5	8.8	25.0	30.0	33.8	11.3	63.8		
Setup 2	0.0	2.5	16.3	37.5	43.8	2.5	81.3		
DV: document content									
Setup 1	8.8	15.0	27.5	17.5	31.3	23.8	48.8		
Setup 2	0.0	2.5	27.5	36.3	33.8	2.5	70.0		
TV: keyword approach									
Setup 1	5.0	22.5	23.8	21.3	27.5	27.5	48.8		
Setup 2	3.8	10.0	16.3	37.5	32.5	13.8	70.0		

(-) ive=scale 1, 2; (+) ive= scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.17: ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups perceived as interesting

This indicates that the participants found the ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features in Setup 2 of *iEvent* were more interesting than Setup 1.

6.5 Detection Task

The entire detection task was successful with 85% of task results being correct and 15% being partially correct. Surprisingly there were no unsuccessful detection tasks nor participants who wrongly detected the topics. This proved that *iEvent* managed to facilitate the participants to perform well in the Detection task as shown in Figure 6.10.

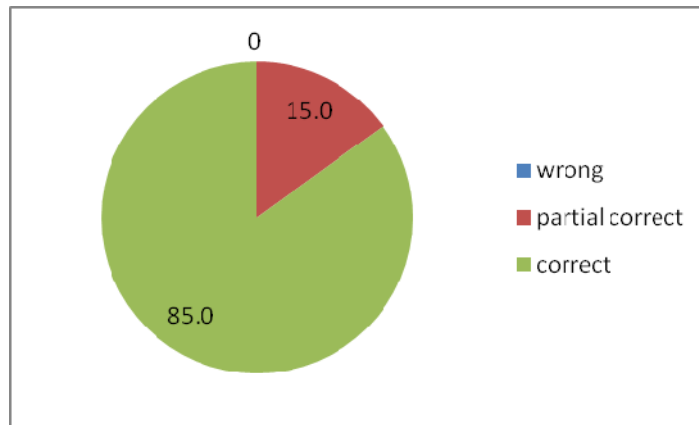


Figure 6.10: Percentage of successful Detection task

A Mann-Whitney Test also confirmed that there was no statistical significance difference in the number of correct topics to be detected ($p=0.534$) in conjunction with the setups. This indicates that the participants managed to detect the correct topics using both setups.

I also classify the correctness of topic detected into four categories:

- i. none- where participants did not provide any information or they did not complete the task
- ii. wrong- where participants detected the wrong topic.
- iii. partially correct-where participants listed out the minor topic as their main finding
- iv. correct-where participants listed out the major topic as their main finding

Interestingly, there were 11 participants to 1 who found that it was easy to detect the topic in this task as shown in Table 6.18. 51.3% found that it was easy to detect the topic (scale 4) and 20% found it was very easy (scale 5) using *iEvent*.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Easy to detect	0.0	6.3	22.5	51.3	20.0	6.3	71.3	11:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.18: Percentage of participants' opinion on *easy to detect*

There was no statistical significance difference on the ease of detecting a topic in conjunction with the cluster given (Mann-Whitney Test, $p=0.735$). This proved that although participants were given a combination of good and poor cluster performance, they manage to complete the Detection task and perform well using *iEvent*. Further results from the interaction logs among the successful tasks showed that participants took 4 minutes and 49 seconds (mean of click=39) to perform this task on average, much less than the 10 minutes given to complete the task.

There was a statistical significance difference between participants opinion on the ease to detect a topic and the setups (Mann-Whitney Test, $p<0.05$). It was easier to detect a topic using Setup 1 than Setup 2 as shown in Figure 7.10. It was apparent from Table 6.19 that 60% of participants agreed that it was easy (scale 4) to detect a topic using Setup 1. Surprisingly none of the participants found that it was hard to detect a topic using Setup 1 and 92.5% found that Setup 1 makes the Detection task easier than Setup 2.

	Scale (%)					(%)	
	1	2	3	4	5	(-)ive	(+)ive
Setup 1	0.0	0.0	7.5	60.0	32.5	0.0	92.5
Setup 2	0.0	12.5	37.5	42.5	7.5	12.5	50.0

(-) ive=scale 1, 2; (+) ive= scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 6.19: Percentage of the participant opinions on the ease to detect a topic across setups

Results also show that there was no statistical significance difference on the ease to detect a topic in conjunction with the type of participant (Mann-Whitney Test, $p=0.477$). Both students and journalists found out that *iEvent* assisted them in detecting the topic easily (mean=3.85 sd=0.813).

The highest percentage of features used was the ‘CV: top terms’ (83.8%) while the lowest was the ‘CV: cluster visualisation’ (53.8%). The participants were using the ‘CV: top terms’ feature to get more information when detecting the topics and less on the ‘CV: cluster visualisation’ feature as shown in Table 6.20. A possible explanation for this might be that participants only deal with a specific assigned cluster without having to compare them with other cluster which makes this feature less useful in the Detection task. Further analyses on the interaction log proved that there was a low activity (2.7%) using the ‘CV: cluster visualisation’ feature.

	Percentage
CLUSTER VIEW (CV)	
cluster labelling	75.0
top terms	83.8
cluster visualisation	53.8
DOCUMENT VIEW (DV)	
histogram with the timeline	80.0
document content	81.3
TERM VIEW (TV)	
keyword approach	81.3
histogram with the time line	57.5

(Higher=better; highest value shown in bold)

Table 6.20: The percentage of *iEvent* features used in Detection task

There were three features which also received a high percentage of use; namely the ‘DV: document content’ (81.3%), ‘TV: keyword approach’ (81.3%) and ‘DV: histogram with the timeline’ (80%) features. Further analyses on the interaction logs proved that there was high activity using the ‘DV: document content’ feature with 77.6% of participants

using it. There was also 3.9% of activity using the ‘TV: keyword approach’ and 7.2% of activity using the ‘DV: histogram with the timeline’.

A possible explanation for this might be that the participants received more information from the ‘DV: document content’ feature to detect the topic and the ‘DV: histogram with the timeline’ feature giving an overall view on the distribution of the topics for the specific cluster. Participants could identify how many topics the cluster contains. While the ‘TV: keyword approach’ feature gives good information on the most frequent terms appearing in the cluster, thus allowing the participant to relate to the topics easier.

I also analysed the frequency of features used during the Detection task across setups. The Mann-Whitney test proved that there was a statistical significance difference on two features across setups for the Detection task. These features were the ‘DV: document content’ ($p=0.008$) and ‘TV: keyword approach’ feature ($p=0.033$) as shown in Table 6.21.

	p-value	Frequency (%)	
		Setup 1	Setup 2
CLUSTER VIEW (CV)			
cluster labelling	1.000	75	75
top terms	0.763	85	83
cluster visualisation	0.824	55	53
DOCUMENT VIEW (DV)			
histogram with the timeline	0.092	98	88
document content	0.008	65	90
TERM VIEW (TV)			
keyword approach	0.033	88	68
histogram with the timeline	1.000	58	58

Table 6.21: The significant value for the frequency of the features across setups (Mann-Whitney Test)

Analysis shows that the participants frequently used ‘DV: document content’ feature (90%) in Setup 2 and ‘TV: keyword approach’ feature (88%) in Setup 1 which helps them to identify the topic.

6.6 Discussion

This chapter described the methodology for *iEvent* evaluation with the main focus on evaluating the features of *iEvent* for each task. The purpose of this experiment was to investigate the effectiveness of *iEvent* (*iTDT* interface) in facilitating the journalists in performing the TDT tasks. I set out to determine which features of *iEvent* facilitate the Tracking and the Detection tasks.

This experiment has shown that generally *iEvent* facilitates the participants to perform well with a high percentage of successful Tracking and Detection tasks. Surprisingly there was only 3.8% of unsuccessful tasks in Tracking and none in Detection task. Findings have revealed that the participants were more familiar with the topics in the Tracking task after using *iEvent*. They were also more interested in the topics in the Tracking task after using Setup 2 of *iEvent*.

	USEFUL	EFFECTIVE	HELPFUL	INTERESTING
CLUSTER VIEW (CV)				
cluster labelling	3:1	2:1	3:1	11:1
top terms	2:1	4:1	6:1	11:1
cluster visualisation	10:1	6:1	6:1	5:1
cluster description button	1:1	1:1	1:1	1:1
DOCUMENT VIEW (DV)				
histogram with the timeline	6:1	7:1	5:1	14:1
document content	5:1	11:1	6:1	5:1
TERM VIEW (TV)				
keyword approach	3:1	5:1	4:1	3:1
histogram with the timeline	6:1	3:1	12:1	5:1

(Higher=better; highest value shown in bold)

Table 6.22: The ratio of each feature across participants’ opinion in the Tracking task

These were the features with the highest ratio that participants perceived as useful, effective, helpful and interesting as shown in Table 6.22. The results revealed that generally Cluster View was useful and interesting Document View was effective and interesting, and Term View was helpful.

For Cluster View 11 participants to 1 agreed the ‘CV: cluster labelling’ and ‘CV: top terms’ features were interesting. The participants found the ‘CV: cluster labelling’ feature in Setup 2 of *iEvent* was more useful, more effective and more interesting than Setup 1. They also found that the ‘CV: top terms’ feature in Setup 2 of *iEvent* was more helpful and more interesting than Setup 1. Meanwhile 10 participants to 1 perceived the ‘CV: cluster visualisation’ feature as useful during the Tracking task, thus this feature received the highest ratio for usefulness. However this feature is the lowest in the Detection task because the participants only deal with one specific cluster to detect the related topics compared to the Tracking task where participants have to track several related clusters. For the Detection task, there was only one feature in Cluster View which received the highest percentage which is the ‘CV: top terms’.

For Document View, 14 participants to 1 found that the ‘DV: histogram with the timeline’ feature was interesting and 11 participants to 1 agreed that the ‘DV: document content’ was effective. Interestingly these two features also received the highest ratio on the opinions mentioned. The participants also found the ‘DV: document content’ feature in Setup 2 of *iEvent* was more interesting than Setup 1. It also appears that the Document View was an important component since two features in it - ‘DV: histogram with the timeline’ and ‘DV: document content’ - received a high percentage in the Detection task. These indicate that the Document View with the features in it does facilitate the participants in performing both tasks. In addition the ‘DV: document content’ feature in Setup 2 of *iEvent* was used more frequently compared to Setup 1 during the Detection task.

For Term View, 5 participants to 1 agreed that the ‘TV: keyword approach’ feature was effective. They also found the ‘TV: keyword approach’ feature in Setup 2 of *iEvent* was more helpful and more interesting than Setup 1. Finally the ‘TV: histogram with the time line’ feature received the highest ratio with 12 participants to 1 agreeing it was helpful. Meanwhile in the Detection task, there was only one feature in Term View that received a high percentage which was the ‘TV: keyword approach’. In addition the ‘TV: keyword approach’ feature in Setup 1 of *iEvent* was used more frequently compared to Setup 2 during the Detection task.

Surprisingly the ‘CV: cluster description button’ feature received the lowest ratio (1:1) across the four opinions since most of the participants perceived its usefulness as average (scale 3). This is supported by the analyses from the interaction logs that indicated a low usage of it during the Tracking task.

	TRACKING	DETECTION
CLUSTER VIEW		
cluster labelling	√	√
top terms	√	√
cluster visualisation	√	×
cluster description button	×	×
DOCUMENT VIEW		
histogram with the timeline	√	√
document content	√	√
TERM VIEW		
keyword approach	√	√
histogram with the timeline	√	×

Table 6.23: The comparison of each feature in facilitating the TDT tasks

As shown in Table 6.23, the Cluster View wit features such as the ‘CV: cluster labelling’ feature and the ‘CV: top terms’ feature facilitate the participants in performing both tasks. Meanwhile the ‘CV: cluster visualisation’ feature only facilitates the participant

during the Tracking task but not for the Detection task due to the nature of the task itself. The Document View with the features in it such as the ‘DV: histogram with the timeline’ and the ‘DV: document content’ does facilitate the participants in performing both tasks. The participants found that the ‘TV: keyword approach’ feature was popular in both tasks. This is because participants need to detect the related topics and ‘TV: keyword approach’ feature allows them to see the most frequent terms in the specific cluster assigned. Meanwhile the ‘TV: histogram with the timeline’ feature was popular during the Tracking task. This probably has to do with the participants’ behaviour in trying to match the pattern of the ‘DV: histogram with the timeline’ feature with the ‘TV: histogram with the timeline’ feature. These results indicate that the ‘TV: keyword approach’ feature facilitates participants in both tasks while the ‘TV: histogram with the timeline’ feature only facilitates participants in the Tracking task.

6.7 Chapter Summary

Overall these findings reveal that the *iEvent* interface generally facilitated the journalists in performing well in the TDT tasks. There were few features in Setup 2 of *iEvent* that facilitated the journalists to perform well in the TDT tasks. This indicates that highlighting the named entities with different colours has affected the participants’ opinions of *iEvent*. Thus it would be interesting to merge Setup 1 and Setup 2 in one interface for the future work on *iEvent*. Therefore journalists have an option to enable the highlighting of named entities in the features of *iEvent*. Some comments were made suggesting revision of the *iEvent* layout which is also interesting for the future work on *iEvent*. In the next chapter I discuss the findings from the evaluation of the use of Named Entity Recognition (NER) in *iEvent*. Chapter 7 is an additional part of the evaluation where I compare the setups of *iEvent*.

Chapter 7

Evaluation of the use of Named Entity Recognition in *iEvent*

7.1 Introduction

It was reported in Chapter 4 that only a few works on interactive Topic Detection and Tracking (*iTDT*) applied Named Entity Recognition (NER) and none of these works had conducted a proper evaluation. This has been the motivation of the work reported in this chapter. Previously I evaluated the usability and the effectiveness of each component of *iEvent* in Chapter 6. I also identified the features of *iEvent* that were effective in performing the TDT tasks. I believe a well designed *iTDT* interface should facilitate the user in performing the TDT tasks better and also create more interaction between user and system. In this work, named entities create context in the interface since this approach is in line with journalists tasks.

In this chapter, I compare two settings of *iEvent*: Setup 1 (baseline setup) and Setup 2 (experimental setup). Setup 1 uses the keywords while Setup 2 uses the named entities. I

investigate the effectiveness of NER in *i*TDT and identify what TDT tasks are facilitated by the use of NER. The aim of this chapter is to evaluate the features of *i*Event with NER in *i*TDT and therefore to answer the following research question; ‘Will the use of NER improve *i*TDT in the same way that it improves standard TDT?’ This is because research in TDT (Allan et al., 1999; Kumaran and Allan, 2004; Li, 2006; Li and Croft, 2005; Makkonen et al., 2004; Otterbacher et al., 2005) have investigated and proved that NER improved TDT system performance.

7.2 Evaluation Methodology

The methodology of the evaluation reported in this chapter used the same methodology as reported in Chapter 3 (Section 3.4). The experimental data used is a selection of 1,468 documents from the TDT2 and TDT3 dataset which comes from the CNN news resources. The participants involved were a combination of journalists and postgraduate journalism students from the Scottish Centre for Journalism Studies (SCJS), University of Strathclyde. They were asked to perform Topic Tracking and Topic Detection task and were provided with the same procedure, training session and the same questionnaires. The difference was that, I compared participants’ performance using different setups of *i*Event and analyse the findings. This was important in order to verify how useful named entities are in helping the journalist to perform the TDT tasks.

The evaluation started with a discussion of the General Findings (Section 7.3), such as the participants’ appreciation of the setups; their topic familiarity and topic interest after using *i*Event across setups; and the participant’s preference between the tasks and setups.

Next the participants’ performance in the Tracking task (Section 7.4) was analysed by examining the setups that participants perceived as easy, relaxing, simple, satisfying and interesting. I also analysed their performance in the Reporting and the Profiling tasks

(sub activities in Tracking) such as the amount of news written, the amount and the types of keywords given to write a profile of a story (terms, named entities or combination of it); and investigated which features of *iEvent* that participants perceived as useful, effective, helpful and interesting, across setups. Then I examined the participants' performance in the Detection task (Section 7.5) such as the participants' opinion on the ease of detecting the topics across setups and the useful features used to perform the Detection task between setups. Finally I compared and explained which TDT tasks are facilitated through the use of NER and discussed the participants' opinions on the features related to the NER.

7.3 General Findings

There are four features of *iEvent* which differentiate between the keywords and the named entities, as shown in Table 7.1.

	Setup 1	Setup 2
CLUSTER VIEW (CV)		
cluster labelling	Keywords	NE
top terms	Keywords	NE
cluster visualisation	×	×
cluster description button	×	×
DOCUMENT VIEW (DV)		
histogram with the timeline	×	×
document content	Keywords	NE
TERM VIEW (TV)		
keyword approach	Keywords	NE
histogram with the time line	×	×

NE=Named Entities

Table 7.1: The differences of the two setups

There is a difference on these four features of *iEvent*; 'CV: cluster labelling', 'CV: top terms', 'DV: document content' and 'TV: keyword approach' between the setups. They

were using named entities in Setup 2 (experimental setup) instead of using keywords in Setup 1 (baseline setup).

Findings revealed that 75% of the participants agreed that Setup 2 helps the journalists' more with the task. Interviews with the participants highlighted that the use of named entities was in line with the journalists' task since it provides significant information on the *Who*, *Where* and *When* of an event. Moreover, 30% of the participants agreed that highlighting the named entities with different colours was the best feature of *iEvent*.

A Mann-Whitney Test confirmed that there was no statistical significance difference ($p=0.492$) in topic interest before using *iEvent* across setups. However there was a statistical significance difference in topic interest after using *iEvent* across setups (Mann-Whitney Test, $p=0.003$). The participants were more interested with a topic in the Tracking task after using Setup 2 (mean=3.81 sd=1.032). Participants found using Setup 2 has enhanced their topic interest (after) as shown in Figure 7.1.

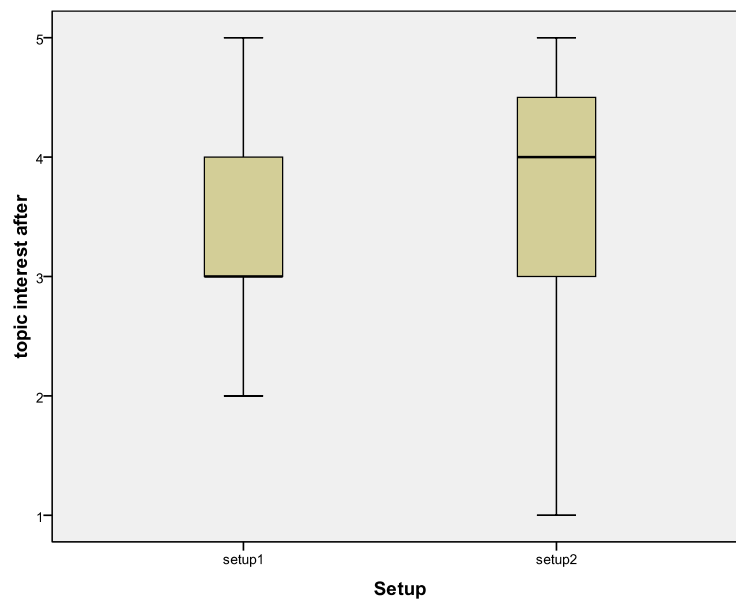


Figure 7.1: Boxplot of participants' topic interest (scale 1-5) after using *iEvent* across setups in the Tracking task

It is apparent from Table 7.2, that there were 7 participants to 1 who found that they were more interested in a topic after using Setup 2. 7 participants to 1 found that the use of NER had significantly enhanced their topic interest with 46.3% of participants agreeing that they were interested (scale 4) with a topic.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	0.0	15.0	37.5	36.3	11.3	15.0	47.5	3:1
Setup 2	5.0	5.0	18.8	46.3	25.0	10.0	71.3	7:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.2: Topic interest (after) across setups

75% of the participants interviewed agreed that their best performance in the experiment was with Setup 2. This indicates that NER helps the participant in performing the TDT tasks. 80% of the participants interviewed noticed the difference between setups where Setup 2 (experimental setup) provided them with named entities.

In the post-evaluation questionnaire, the participants were asked to circle the setup that they felt was useful in performing the Tracking, Reporting, Profiling and Detection tasks. The statistics revealed that the use of NER helped to facilitate the participant in the Tracking, Reporting and Profiling tasks while keywords help in the Detection task, as shown in Figure 7.2. During the post-evaluation interview, the participants informed me that the named entities provided them with high quality forms of information while keywords were more descriptive. Thus Setup 2 was significant in the Tracking task (Reporting and Profiling) where the participants required specific and meaningful information. Setup 1 was helpful in the Detection task because it provided participants with broad information. Participants need more information when they want to detect the topics in the Detection task and Setup 1 provides them with all the keywords.

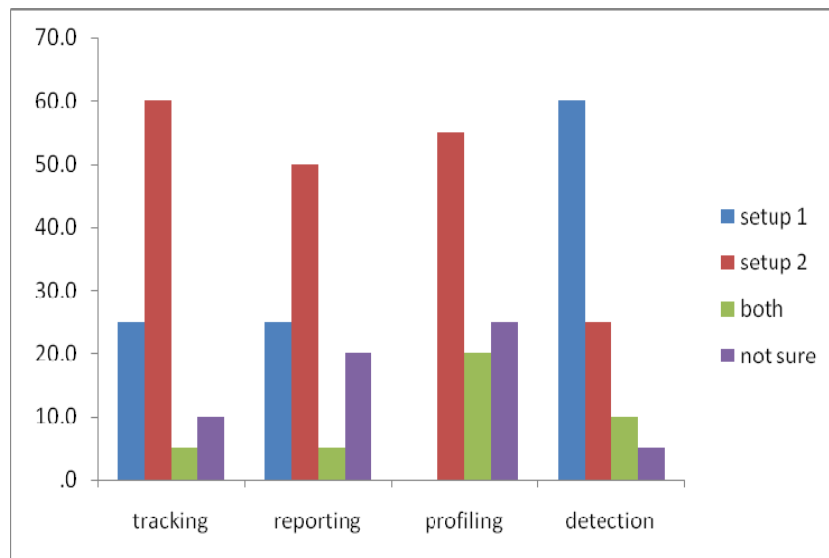


Figure 7.2: The percentage of the participants' preferred setup for the given tasks

7.4 Tracking Task

Several analyses were performed on the captured data. The following sections present the findings. First, the participants' overall opinions on the use of NER in Setup 2 were examined. Next, I investigate the participants' performance in the Reporting task such as the amount of news written and in the Profiling task such as the amount and the types of keywords given to write a profile of a story (terms, named entities or combination of it). Finally I investigated whether the participants agreed that the use of NER in *iEvent* was perceived as useful, effective, helpful and interesting in performing the TDT tasks.

7.4.1 Overall Opinions

Participants' opinions of *iEvent* during the Tracking task were analysed between setups. I investigated whether they perceived *iEvent* as easy, relaxing, simple, satisfying and interesting across setups. Mann-Whitney Test confirmed that there was no statistical significance difference in participants' opinion in simple ($p=0.840$) and satisfying

($p=0.500$) in conjunction with the setups. However there was a statistical significance difference in easy ($p=0.004$), relaxing ($p=0.003$) and interesting ($p<0.05$). Setup 2 was significantly easier, relaxing and interesting (scale 4) compared to Setup 1 as shown in Figure 7.3.

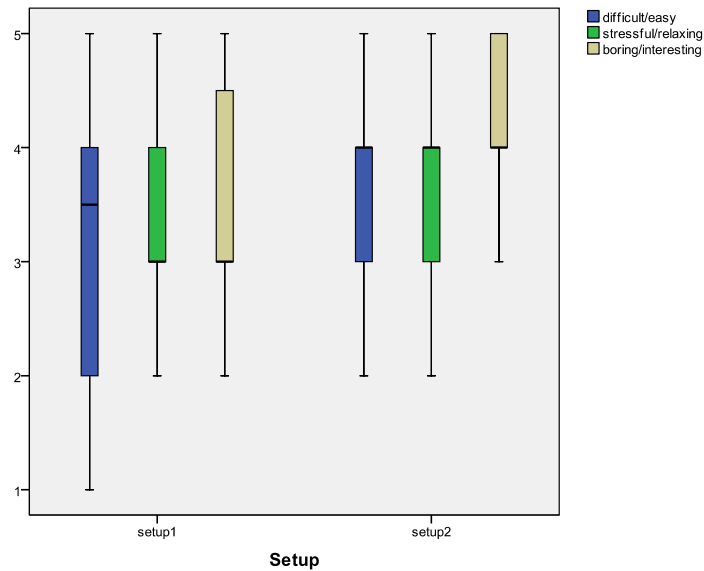


Figure 7.3: Boxplot of participants' opinion (scale 1-5) for the setups in the Tracking task

I analysed Setup 2 of *iEvent* that participants perceived as easier, relaxing and interesting during the Tracking task as shown in Table 7.3.

Opinion	Scale (%)					(%)	
	1	2	3	4	5	(-)ive	(+)ive
Easy	0.0	5.0	27.5	45.0	22.5	5.0	67.5
Relaxing	0.0	7.5	32.5	41.3	18.8	7.5	60.0
Interesting	0.0	0.0	21.3	35.0	43.8	0.0	78.8

(-) ive=scale 1, 2; (+) ive= scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.3: Percentage of the participant opinions of Setup 2

a. Easy

45.0% of participants agreed that Setup 2 (mean=3.85 sd=0.828) was easy (scale 4). Interestingly there were 67.5% of participants who found that Setup 2 was easier compared to 5% which found it difficult. This indicates that 14 participants to 1 found that the use of NER made the Tracking task easier.

b. Relaxing

41.3% of participants agreed that Setup 2 (mean=3.71 sd=0.860) was relaxing (scale 4). There were 60% of participants who found that Setup 2 was more relaxing compared to 7.5% which found it stressful.

c. Interesting

Setup 2 (mean=4.23 sd=0.779) received the highest percentage with 43.8% of participants agreeing that it was very interesting (scale 5). Surprisingly none of the participants found that Setup 2 was boring and 78.8% of participants found that it was more interesting.

During the interview session, participants agreed that use of NER gave them quick and precise information on the *Who, Where, When* of an event that made the Tracking task easy and relaxing using Setup 2. According to them, Setup 2 was also interesting because it highlighted named entities; especially in the 'DV: document content' feature. Further analyses on the interaction logs proved that the participants had a higher activity of this feature in Setup 2 (74.3%) compared to Setup 1 (68.6%).

The average time taken in performing the Tracking task successfully was almost the same using both approaches. Participants took an average of 13 minutes 57 seconds using named entities (Setup 2) and 14 minutes 24 seconds using the keywords (Setup 1).

7.4.2 Reporting Task

This section reports the findings of participants' performance during the Reporting task as one of the sub activities in Tracking. I analysed the number of lines that participants wrote across setups. There was no statistical significance difference on the amount of news written in conjunction with the setups (Mann-Whitney Test, $p=0.434$). The participants managed to write the amount of news equally using Setup 1 (mean=9.15 sd=6.611) and Setup 2 (mean=9.73 sd=6.325).

7.4.3 Profiling Task

The participants were required to provide the important keywords as a profile for a topic. There are three types of keywords; named entities, terms and combination (terms and named entities). For example, participants might provide keywords such as *Oprah* (named entity), *mad cow disease* (terms) and *Oprah lawsuit* (combination) for topic *Oprah Lawsuit*. I analyse the frequency of the types of keywords provided by the participants.

The participant provides a large number of named entities (46.6%) as keywords used in the Profiling task. There were 875 keywords collected during the Tracking task as shown in Figure 7.4. Participants provided an average of 5 keywords for each topic.

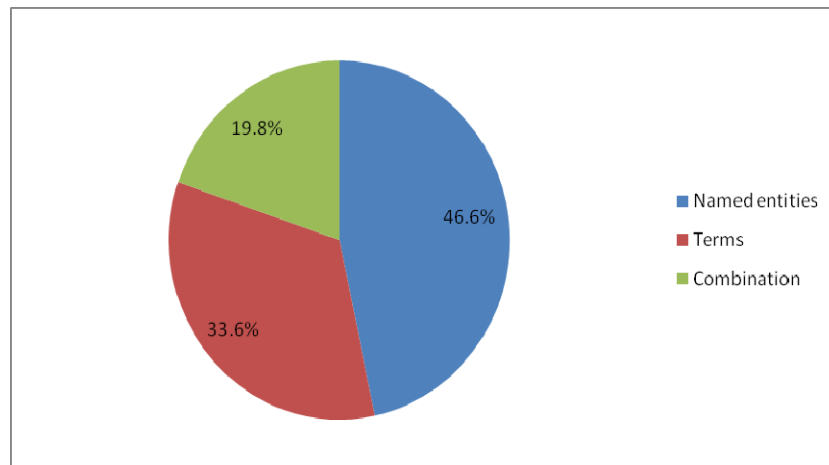


Figure 7.4: Percentage for the type of keywords

There was no statistical significance difference on the types of keywords used in conjunction with the topics in the Profiling task (Mann-Whitney Test, $p=0.912$). Participants provided a balanced amount of named entities, terms and combination as the important keywords for most topics during the Profiling tasks, as shown in Table 7.4. The most interesting results were that the participants provided the highest amount of keywords ($n=135$) for the topic *Cable Car Crash* and they provided a large number of named entities ($n=81$) for the topic *October Holbrooke-Milosevic Meeting*.

Topics	Named entities	Terms	Combination	Total
Oprah Lawsuit	51	40	23	114
National Tobacco Settlement	53	42	15	110
Cable Car Crash	51	50	34	135
Jonesboro shooting	44	40	21	105
October Holbrooke-Milosevic Meeting	81	10	21	112
Mobil-Exxon Merger	47	39	18	104
German Train derail	36	40	21	97
Asteroid Coming??	45	33	20	98
	408	294	173	875

Table 7.4: The frequency for types of keywords across topics

There was a statistically significant difference between the types of keywords and the setups used in the Profiling task (Mann-Whitney test, $p=0.002$). The participants provided a large number of named entities (52%) when they were using Setup 2. It seems that the participants also provided a large number of named entities (41%) when they were using Setup 1, as shown in Figure 7.5.

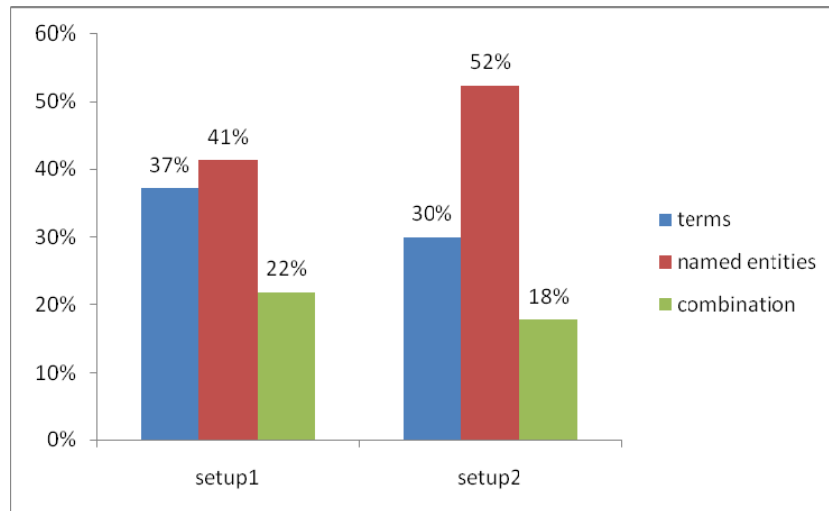


Figure 7.5: The percentage for the types of keywords across setups

This revealed that participants preferred named entities as the important keywords when they wanted to write a profile of a story, compared with using terms. The participants interviewed agreed that named entities were a high quality source of information, thus they were suitable to be used as the profile keywords rather than using terms in the Profiling task. I therefore conclude that named entities lead to a better topic profile. It seems that participants need a way for the interface to automatically provide them with the named entities rather than them having to identify it by themselves. Thus a useful *i*TDT interface should support a means for automatically providing the users with the named entities.

7.4.4 Features

In this section, I analyse each feature of *iEvent* across setups that participants perceived as being a useful, effective, helpful and interesting during the Tracking task.

a. Useful

For this opinion, there was a statistical significance difference on the ‘CV: cluster labelling’ feature across setups (Mann-Whitney Test, $p < 0.05$). Setup 2 was more useful (mean=3.94 sd=0.919) compared to Setup 1 (mean=3.16 sd=1.267). Participants found this feature was useful when they were using Setup 2 as shown in Figure 7.6.

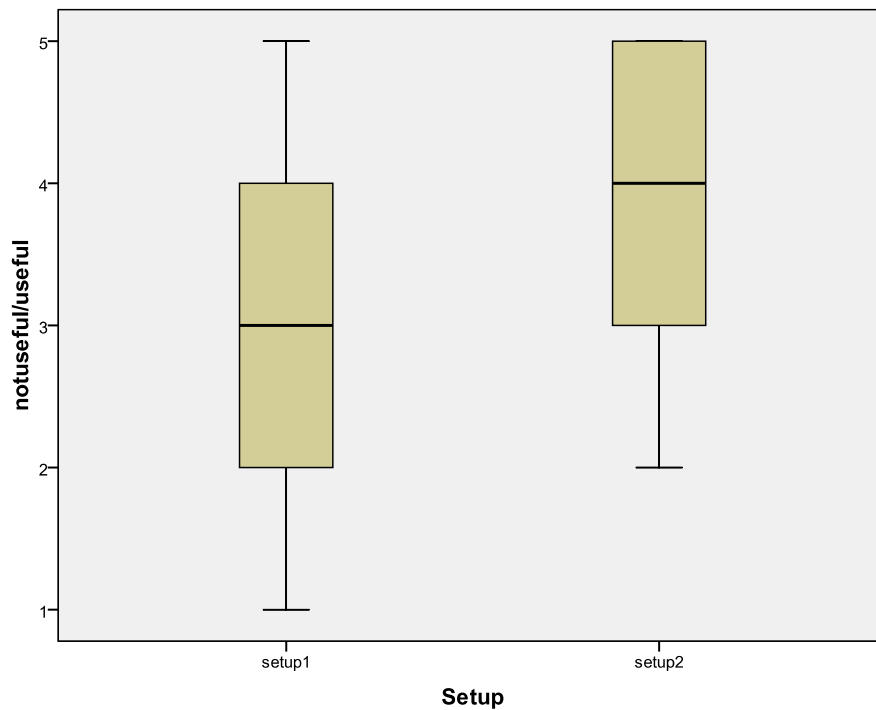


Figure 7.6: Boxplot on the *usefulness* (scale 1-5) of the ‘CV: cluster labelling’ feature across setups

9 participants to 1 found that NER used in ‘CV: cluster labelling’ feature were perceived as significantly useful with 38.8% of participants agreed that it was useful (scale 4) as shown in Table 7.5.

	Scale (%)					(-)ive (%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	11.3	23.8	17.5	32.5	15.0	35.0	47.5	1:1
Setup 2	0.0	7.5	22.5	38.8	31.3	7.5	70.0	9:1

(-) ive= scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.5: ‘CV: cluster labelling’ feature across setups perceived as useful

b. Effective

For this opinion, again there was a statistical significance difference on the ‘CV: cluster labelling’ feature across setups (Mann-Whitney Test, p=0.008). Setup 2 was more effective (mean=3.66 sd=0.927) compared to Setup 1 (mean=3.21 sd=1.110). Participants found this feature was effective when they were using Setup 2 as shown in Figure 7.7.

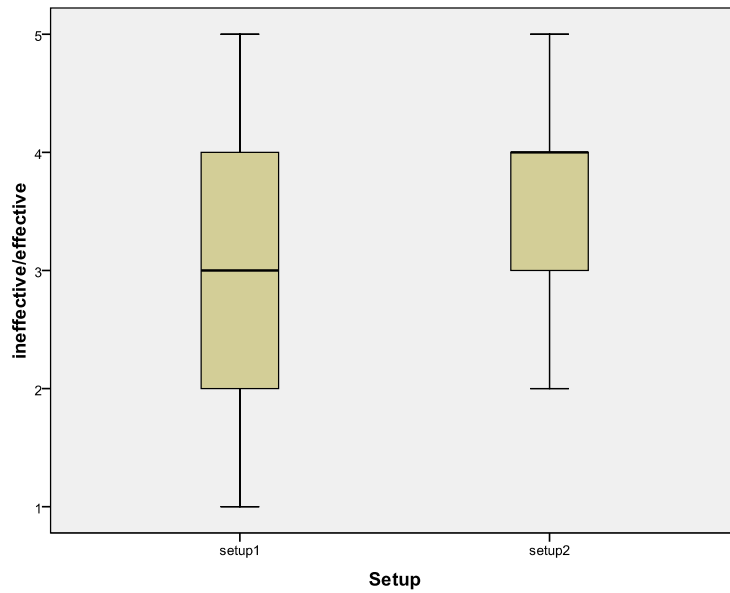


Figure 7.7: Boxplot on the *effectiveness* (scale 1-5) of the ‘CV: cluster labelling’ feature across setups

5 participants to 1 found that NER used in ‘CV: cluster labelling’ feature was perceived as significantly effective with 41.3% of participants agreeing that it was effective (scale 4) as shown in Table 7.6.

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
Setup 1	3.8	26.3	30.0	25.0	15.0	30.0	40.0	1:1
Setup 2	0.0	12.5	27.5	41.3	18.8	12.5	60.0	5:1

(-) ive=scale 1, 2; (+) ive=scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.6: ‘CV: cluster labelling’ feature across setups perceived as effective

c. Helpful

For this opinion, there was a statistically significant difference on two features across setups. These were the ‘CV: top terms’ feature (Mann-Whitney Test, $p=0.033$) and ‘TV: keyword approach’ feature (Mann-Whitney Test, $p=0.011$).

For the ‘CV: top terms’ feature, Setup 2 was more helpful (mean=3.95 sd=0.913) compared to Setup 1 (mean=3.59 sd=1.040). Meanwhile for the ‘TV: keyword approach’ feature, Setup 2 was also more helpful (mean=3.95 sd=0.940) compared to Setup 1 (mean=3.51 sd=1.067). Participants found these two features were helpful when they were using Setup 2 as shown in Figure 7.8.

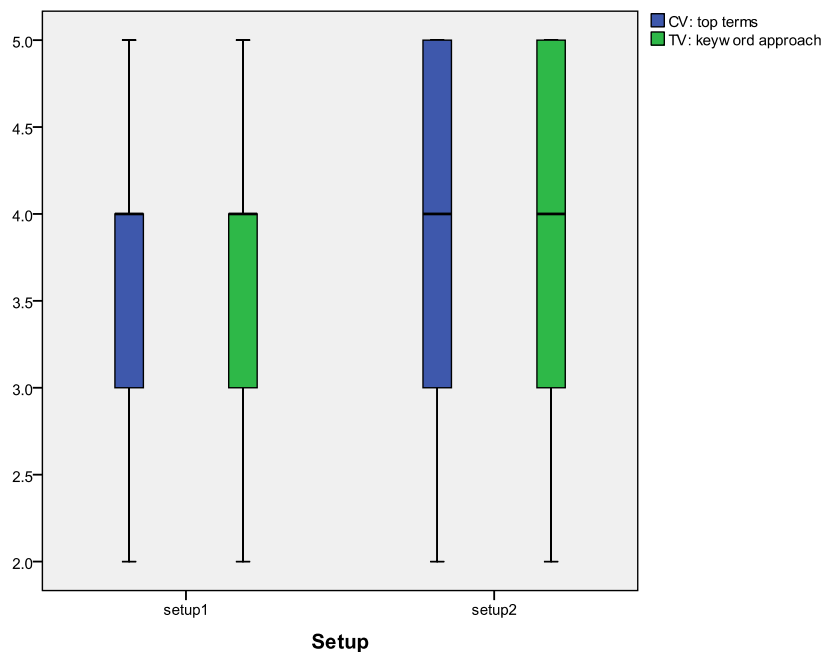


Figure 7.8: Boxplot on the *helpfulness* (scale 1-5) of the ‘CV: top terms’ feature and ‘TV: keyword approach’ feature across setups

Surprisingly 17 participants to 1 found that NER used in ‘CV: top terms’ feature was perceived as significantly helpful with 35.0% of participants agreeing that it was very helpful (scale 5) as shown in Table 7.7. 9 participants to 1 found that NER used in ‘TV:

keyword approach’ feature was perceived as significantly helpful with 35.0% of participants agreeing that it was helpful (scale 4).

	Scale (%)					(%)		Ratio
	1	2	3	4	5	(-)ive	(+)ive	
CV: top terms								
Setup 1	1.3	15.0	30.0	31.3	22.5	16.3	53.8	3:1
Setup 2	0.0	3.8	32.5	28.8	35.0	3.8	63.8	17:1
TV: keyword approach								
Setup 1	1.3	22.5	17.5	41.3	17.5	23.8	58.8	2:1
Setup 2	0.0	7.5	23.8	35.0	33.8	7.5	68.8	9:1

(-) ive=scale 1, 2; (+) ive= scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.7: ‘CV: top terms’ and ‘TV: keyword approach’ features across setups

d. Interesting

For this opinion, there was a statistically significant difference on four features across setups. These were the ‘CV: cluster labelling’ feature (Mann-Whitney Test, p=0.033), ‘CV: top terms’ feature (Mann-Whitney Test, p=0.026), ‘DV: document content’ feature (Mann-Whitney Test, p=0.013) and ‘TV: keyword approach’ feature (Mann-Whitney Test, p=0.035).

Interestingly these four features of *iEvent* in Setup 2 such as ‘CV: cluster labelling’ (mean=4.29 sd=0.814), ‘CV: top terms’ (mean=4.23 sd=0.811), ‘DV: document content’ (mean=4.01 sd=0.849) and ‘TV: keyword approach’ (mean=3.85 sd=1.104) were perceived as being interesting compared to Setup 1 by the participants as shown in Figure 7.9.

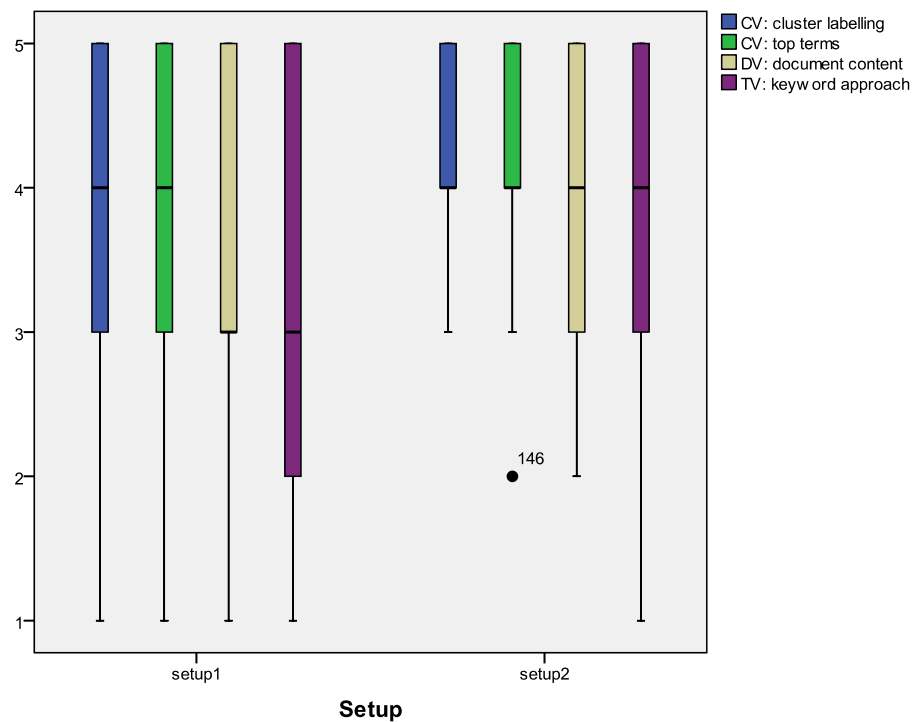


Figure 7.9: Boxplot on the *interestingness* (scale 1-5) of the ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups

Table 7.8 shows that there were two features in Setup 2 that were perceived as very interesting (scale 5) by the participants. These were the ‘CV: cluster labelling’ feature (48.8%) and the ‘CV: top terms’ feature (43.8%). There were also two further features in Setup 2 perceived as interesting (scale 4) by the participants; the ‘DV: document content’ feature (36.3%) and the ‘TV: keyword approach’ feature (37.5%). Surprisingly these four features received a high percentage (70%-82.5%) of participants who found it interesting. It was clear that participants found that the use of NER was more interesting than the keywords in the Tracking task.

	Scale (%)					(-)ive (%)		(+)ive (%)	
	1	2	3	4	5				
CV: cluster labelling									
Setup 1	1.3	8.8	36.3	17.5	36.3	10.0		53.8	
Setup 2	0.0	2.5	15.0	33.8	48.8	2.5		82.5	
CV: top terms									
Setup 1	2.5	8.8	25.0	30.0	33.8	11.3		63.8	
Setup 2	0.0	2.5	16.3	37.5	43.8	2.5		81.3	
DV: document content									
Setup 1	8.8	15.0	27.5	17.5	31.3	23.8		48.8	
Setup 2	0.0	2.5	27.5	36.3	33.8	2.5		70.0	
TV: keyword approach									
Setup 1	5.0	22.5	23.8	21.3	27.5	27.5		48.8	
Setup 2	3.8	10.0	16.3	37.5	32.5	13.8		70.0	

(-) ive=scale 1, 2; (+) ive= scale 4, 5

(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.8: ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’ features across setups

7.5 Detection Task

This section is concerned with the analysis of participants’ agreement on the ease of detecting the topics and the frequency of features with named entities across setups.

There was a statistical significance difference between participants’ opinion on the ease of detecting a topic and the setups (Mann-Whitney Test, $p < 0.05$). It was easier to detect a topic using Setup 1 than Setup 2 as shown in Figure 7.10.

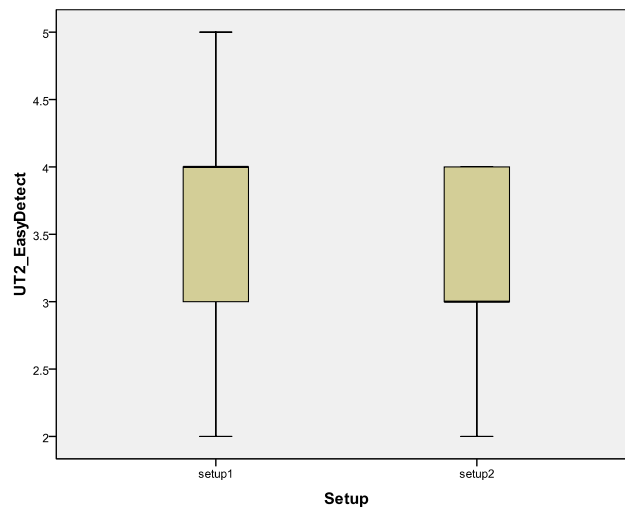


Figure 7.10: Boxplot on *Easy to Detect* (Scale 1-5) of the setup for the Detection task

Analysis shows that the keywords (Setup 1) make the Detection task easier (mean=4.25 sd=0.588) than named entities. As mentioned previously in Section 7.3 (General Findings), during the post-evaluation interview, the participants found that the keywords were more descriptive because the participants had to detect the topic dealt by a specific cluster in the Detection task. They require not only the *Who*, *Where* and *When* but also the *What* information. Thus the keywords support the Detection task because it can explain better the *What*; that is very difficult to capture with NER.

It was apparent from Table 7.9 that 60% of participants agreed that it was easy (scale 4) to detect a topic using Setup 1. Surprisingly none of the participants found that it was hard to detect a topic using Setup 1 and 92.5% found that Setup 1 makes the Detection task easier than Setup 2.

	Scale (%)					(%)	
	1	2	3	4	5	(-)ive	(+)ive
Setup 1	0.0	0.0	7.5	60.0	32.5	0.0	92.5
Setup 2	0.0	12.5	37.5	42.5	7.5	12.5	50.0

(-) ive=scale 1, 2; (+) ive= scale 4, 5
(scale from 1 to 5, higher=better; highest value shown in bold)

Table 7.9: Percentage of the participant opinions on the ease to detect a topic across setups

The Mann-Whitney test proved that there was a statistical significance difference on two features with NER across setups for the Detection task. These features were the ‘DV: document content’ (p=0.008) and ‘TV: keyword approach’ feature (p=0.033) as shown in Table 7.10.

	p-value	Frequency (%)	
		Setup 1	Setup 2
CLUSTER VIEW (CV)			
cluster labelling	1.000	75	75
top terms	0.763	85	83
cluster visualisation	0.824	55	53
DOCUMENT VIEW (DV)			
histogram with the timeline	0.092	98	88
document content	0.008	65	90
TERM VIEW (TV)			
keyword approach	0.033	88	68
histogram with the timeline	1.000	58	58

Table 7.10: The significant value for the frequency of the features with NER across setups (Mann-Whitney Test)

Analysis shows that the participants frequently used ‘DV: document content’ feature (90%) in Setup 2 and ‘TV: keyword approach’ feature (88%) in Setup 1 which helps them to identify the topic.

A possible explanation for these results might be that participants require all keywords to occur in the specific cluster in order to detect the topic rather than presenting only the *Who*, *Where* and *When*. Further analyses on the interaction logs proved that the participants have a higher activity of the ‘TV: keyword approach’ feature in Setup 1 (4.8%) compared to Setup 2 (3.0%). Analyses on the interaction logs also proved that the participants have a slightly higher activity of ‘DV: document content’ feature in Setup 1 (78.7%) compared to Setup 2 (76.6%). These findings revealed that participants clicked less on the ‘DV: document content’ feature with named entities (Setup 2) than with Setup 1. It seems possible that providing them with named entities by highlighting them in the ‘DV: document content’ feature appeared to be a shortcut since this approach provides them with quick and significant information rather than displaying plain document content. It also appeared that among the successful attempts to the Detection task, the participants took less time using named entities (00:04:36) compared to using keywords (00:05:53).

These results indicate that the keywords used in the ‘TV: keyword approach’ feature provide them with all the terms and the named entities used in ‘DV: document content’ feature helps them to make a quick decision of a topic in the Detection task.

7.6 Discussion

Generally participants found that most of the features with NER facilitate them in performing the Tracking task while keywords facilitate them in performing the Detection task. However participants found that the use of NER in ‘DV: document content’ feature helps them to detect the topic quickly. Participants found that they received broad information using keywords; however, the use of NER allowed them to detect the topics faster. This proved that keywords were effective and the use of NER was efficient in the Detection task. Keywords resemble the *What* category and provide a better explanation in the Detection task which requires the participants to gather as much information as possible. Meanwhile capturing the named entities provides quick

information for the participants to decide on the topics without having them to dwell on the whole document content. Combination of these two approaches will help the participants to perform better in the Detection task. Table 7.11 shows the features with NER which facilitate the participants in performing the TDT tasks.

	TRACKING	DETECTION
CLUSTER VIEW (CV)		
cluster labelling	NE	Keywords
top terms	NE	Keywords
DOCUMENT VIEW (DV)		
document content	NE	NE
TERM VIEW (TV)		
keyword approach	NE	Keywords

NE=Named Entities

Table 7.11: The comparison of features with approaches in facilitating the TDT tasks

The Mann-Whitney test proved that there was no statistical significance difference between the successful Tracking ($p=0.160$) and Detection task ($p=0.534$) across setups. The use of NER does not influence the participants to either track the correct cluster or to detect the correct topics better. The participants agreed that keywords help them to detect the topic easier. Although the use of NER does not influence the participants to track the correct cluster, participants did agree that some of the features with NER were useful, effective, helpful and surprisingly, most of them agreed that it was interesting. Thus it could be suggested that the *interesting* aspect led to an increase in the interaction between the participants and the system while performing the Tracking task.

Table 7.12 shows the comparison of the mean numbers of click among the successful tasks between setups as analysed from the interaction log.

Task	Mean of click	
	Setup 1	Setup 2
Tracking	32	45
Detection	43	36

(highest value shown in bold)

Table 7.12: Comparison on the mean of click for each successful task between setups

The mean of clicks using Setup 2 was higher in the Tracking task but lower in the Detection task. When the participant was using Setup 2, they clicked more during the Tracking task but clicked less during the Detection task. The high number of clicks did not indicate that they were lost since the analysis was on the successful tasks. These findings showed that the highlighting of named entities enhances the interaction between the participant and the system in the Tracking task.

The use of NER also leads to a better topic profile. It was an advantage for an *i*TDT interface that supports the users with named entities since they are significant and present a high quality source of information.

7.7 Guidelines for the designs of the *i*TDT interfaces

This section discusses the guideline for the design of the *i*TDT interfaces as depicted in Table 7.13.

	GUIDELINE			
	TRACKING		DETECTION	
	Feature	Approach	Feature	Approach
CLUSTER VIEW (CV)				
cluster labelling	√	NE	√	Keywords
top terms	√	NE	√	Keywords
cluster visualisation	√	-	×	-
cluster description button	×	-	×	-
DOCUMENT VIEW (DV)				
histogram with the timeline	√	-	√	-
document content	√	NE	√	NE
TERM VIEW (TV)				
keyword approach	√	NE	√	Keywords
histogram with the timeline	√	-	×	-

NE=Named Entities

Table 7.13: Guideline for the designs of the *i*TDT interfaces

The Cluster View with the features in it such as the ‘CV: cluster labelling’ feature and the ‘CV: top terms’ feature facilitate journalists in performing both tasks. It would be effective and interactive to provide these features with named entities in the Tracking task. Furthermore journalists found the ‘CV: cluster labelling’ feature in Setup 2 of *i*Event more useful, more effective and more interesting than Setup 1. They also found that the ‘CV: top terms’ feature in Setup 2 of *i*Event was more helpful and more interesting than Setup 1. Meanwhile it was effective and interactive to provide these features with keywords in the Detection task. This is because journalists require significant and a high quality forms of information such as the Who, Where and When, when performing the Tracking task; meanwhile they require broad information when performing the Detection task. The ‘CV: cluster visualisation’ feature only facilitates journalists during the Tracking task but not for the Detection task. They could use the information such as the size and the density of the cluster to track the correct cluster in the Tracking task while they only dealt with one specific cluster in the Detection task. Therefore using ‘CV: cluster visualisation’ feature is an advantage in the Tracking task.

For example they would aim for clusters with large size and high density which might indicate the cluster contained an important event rather than to go for cluster with low size and low density.

The Document View with the features in it such as the ‘DV: histogram with the timeline’ and the ‘DV: document content’ does facilitate journalists in performing both tasks. It would be effective and interactive to provide ‘DV: document content’ feature with named entities in the Tracking task since they require specific information. Furthermore the participants found the ‘DV: document content’ feature in Setup 2 of *iEvent* more interesting than Setup 1. It was also efficient to provide them with named entities in this feature during the Detection task since they could easily detect the topic. This is the reason why journalists used the ‘DV: document content’ feature in Setup 2 of *iEvent* more frequently as compared to Setup 1 during the Detection task.

Finally the ‘TV: keyword approach’ feature does facilitate journalists in performing both tasks. It would be effective and interactive to provide ‘TV: keyword approach’ feature with named entities in the Tracking task. Furthermore they also found the ‘TV: keyword approach’ feature in Setup 2 of *iEvent* more helpful and more interesting than Setup 1. Meanwhile it was effective and interactive to provide ‘TV: keyword approach’ feature with keywords in the Detection task. Tracking task requires them to identify the correct cluster and the use of named entities is an advantage because it has narrowed their searching by providing them with significant terms. Meanwhile Detection task requires them to detect the topic and the use of keywords is an advantage because it has broadened their searching by providing with all terms. This is the reason why journalists used the ‘TV: keyword approach’ feature in Setup 1 of *iEvent* more frequently compared to Setup 2 during the Detection task.

The ‘TV: histogram with the timeline’ feature only facilitates journalists during the Tracking task but not for the Detection task. Journalists used this feature to confirm

whether they had tracked the correct cluster by matching the pattern of the ‘DV: histogram with the timeline’ feature with the ‘TV: histogram with the timeline’ feature.

7.8 Chapter Summary

In conclusion, the use of NER was effective and created more interaction between participants and the system in the Tracking task. The participants took less time using Setup 2 in the Detection task, with the time difference between setups being 1 minute 17 seconds. This proved that NER was efficient in the Detection task since it helped the participants to detect the topic fast. However none of these approaches was efficient during the Tracking task since the time taken between setups were almost the same, and the time difference was 27 seconds. Table 7.14 compared which approach was effective, interactive and efficient in the TDT tasks.

	TRACKING	DETECTION
Effective	NE	Keywords
Interactive	NE	Keywords
Efficient	-	NE

NE=Named Entities

Table 7.14: Comparison of the approaches across TDT tasks

There was a high demand for the *Who*, *Where* and *When* by the participants when performing the Tracking task. They required specific information about a topic, thus the use of NER was effective. As a result, they had a high amount of interaction using the named entities in performing the Tracking task.

Meanwhile keywords were effective and created more interaction between participants and the system during the Detection task. This is because participants require broad information that is equivalent to *What*. Thus they have a high interaction using keywords in performing the Detection task. However the use of NER is efficient in the Detection

task because it helps the participants to detect a topic fast. The named entities allow the participants to skim over the document and make a quick decision about a topic compared to keywords.

The journalists provide a large number of named entities during the Profiling task using both setups. This indicates the importance of providing them with named entities on the interface. Therefore an *i*TDT interface with NER seems to be useful in helping the journalists to create a better topic profile.

In the next chapter I discuss and conclude the findings from Chapter 6 and Chapter 7.

Chapter 8

Conclusion and Future Work

8.1 Introduction

In this thesis I investigated the best features of interactive Topic Detection and Tracking (*i*TDT) interfaces and implemented these into the *i*Event (Interactive Event Tracking System) interface. I have also evaluated the features introduced and established a set of guidelines for future *i*TDT interface design.

*i*Event is aimed at facilitating professionals such as journalists or information analysts to perform TDT tasks. Interestingly this is the first TDT work which involves ‘the journalist’ as a user of an *i*TDT interface. *i*Event is composed of three components which are the Cluster View (CV), Document View (DV) and Term View (TV) and it has two settings which are keywords (baseline setup) and named entities (experimental setup). I also evaluated user opinion on the usability of the features introduced in *i*Event and investigated the effect of *i*Event on the effectiveness of the user performance. This is important in verifying which features of *i*Event facilitate the user to perform TDT tasks.

In addition I also look at how useful Named Entity Recognition (NER) in *i*TDT is by describing which features and TDT tasks are facilitated through the use of it. Finally in this chapter I conclude the thesis by summarising the main contributions of the work. I also propose some opportunities for future research.

The three main objectives of this thesis, as outlined in Section 1.3, were:

- a. Presenting the design of an *i*TDT interface which facilitates journalists to perform the TDT tasks.
- b. Evaluating the *i*Event interface by identifying its useful components. In addition, investigating which features of *i*Event facilitate the Tracking and the Detection task.
- c. Evaluating the use of NER by comparing two different setups of *i*Event interface. In addition, investigating which TDT tasks are facilitated through the use of NER.

These objectives form the structure of this chapter. First, in Section 8.2, I outline the contributions made with respect to each objective. Then, in Section 8.3, I propose how the research associated with each objective could be extended to further benefit the *i*TDT research community.

8.2 Contributions

In summary, the work reported in this thesis made a number of original contributions, which will be reported in what follows.

8.2.1 Novel *i*TDT Interface

A key contribution of this thesis was the design of a novel *i*TDT interface. I designed an *i*Event interface that incorporated a number of successful components and features from existing *i*TDT interfaces into a single interface. Generally *i*Event facilitates the participants in performing well with a high percentage of successful Tracking and Detection tasks. As a result, guidelines for the designs of *i*TDT interfaces have emerged as summarised in Table 7.13 (Section 7.7).

The features such as ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: histogram with the timeline’, ‘DV: document content’ and ‘TV: keyword approach’ facilitates users in performing both tasks. Meanwhile the ‘CV: cluster visualisation’ feature and ‘TV: histogram with the timeline’ only facilitate users during the Tracking task.

NER facilitates users in performing the Tracking task. The use of NER was effective in features such as the ‘CV: cluster labelling’, ‘CV: top terms’, ‘DV: document content’ and ‘TV: keyword approach’. This is because users require specific information on the *Who*, *Where* and *When* in performing the Tracking task. Meanwhile keywords facilitate users in performing the Detection task. Keywords were effective in features such as the ‘CV: cluster labelling’, ‘CV: top terms’ and ‘TV: keyword approach’. Users require broad information, thus using keywords offered a better explanation in the Detection task as it resembles the *What*. However participants found that the use of NER was efficient in the ‘DV: document content’ feature since it helps them to detect the topic fast during the Detection task.

The guideline shall contribute to the effectiveness of users’ performance in performing the TDT tasks. Therefore, it is important for the TDT interface designer to consider the guidelines for the designs of an *i*TDT interface as reported in this thesis.

8.2.2 Proper Evaluation on *i*TDT

Another contribution of this thesis was the creation of a methodology for proper evaluation of *i*TDT. The focus of the methodology is the user tasks. They were designed to follow naturalistic news monitoring behaviour and to support journalists in performing TDT tasks in a way it is in line with their job. It is this feature that makes the current work a ground-breaking study that investigates ‘journalists’ as *i*TDT interface users. In the Tracking task, the participants had to track the cluster that contains the given topic and in the Detection task, they had to identify the topic dealt with by a specific cluster. The tasks aimed to guide other researchers when evaluating *i*TDT user interfaces and aspects of user interaction.

8.2.3 The Importance of Named Entity Recognition (NER) in *i*TDT

The use of NER improves *i*TDT in the same way that it improves standard TDT. Findings revealed that the use of NER is effective, efficient and creates a high level of interaction between the user and the system in *i*TDT. The comparison of two settings of *i*Event revealed that NER is more useful in the Tracking task, since it was effective and enhanced user interaction with the system. It was also efficient in the Detection task. Therefore the use of NER in *i*TDT is useful in helping journalists to perform TDT tasks.

8.3 Future Works

The previous section summarised the main contributions of this work, the implications of which are very important. A number of opportunities for further research have emerged and will be discussed in greater detail in the sections that follow.

8.3.1 Novel *i*TDT Interface

This thesis has highlighted the importance of an *i*TDT interface by identifying the guidelines based on TDT tasks. However, this work is limited to the design of an *i*TDT interface and more work is required to improve the layout issues of an *i*TDT interface. The current layout requires the 21” monitor for best interface display rather than a smaller monitor. Moreover rather than displaying each component of *i*Event (Cluster View, Document View and Term View) in a horizontal way, experiments could be devised to display the components in a vertical way or using other layout techniques. It would be interesting to discover the impact of the layout on the effectiveness of system performance and also user opinion in performing the TDT tasks.

There is also scope for additional research into graphical visualisation techniques in *i*Event interfaces. Besides using cluster visualisation and colour highlighting, techniques such as *brushing and linking* (Eick and Wills, 1995; Hearst, 1995; Shneiderman & Aris, 2006) and *panning and zooming* (Bederson et al., 1996; Bederson, 2001; Hearst, 2009) could be useful too. These techniques support dynamic and interactive use. *Brushing and linking* is useful in connecting the documents in Document View and the terms in Term View for the same cluster. Meanwhile the *panning and zooming* technique could allow the users to zoom the clusters and show the documents or the terms associated with an individual document. These techniques help to make the main view of *i*Event, which is Cluster View, become the users’ main focus of attention, thus enabling the layout of *i*Event to be displayed on a small screen. It would also be interesting to apply personalization techniques (Vallet et al., 2006; Carman and Crestani, 2008) where users could store their interest and interaction with the system; for example, if they are more interested with the *Who* when monitoring the sports news.

8.3.2 Proper Evaluation on *i*TDT

The main focus of the evaluation of the *i*TDT interface was on the Tracking task which is more comprehensive than the Detection task. The evaluation of the *i*TDT interface on the Tracking task has investigated the usability of features introduced in *i*Event. This includes participants' opinion on each feature of *i*Event and NER in facilitating them to perform TDT tasks. Meanwhile, for the Detection task, the evaluation did not include participants' opinion on the usability of each feature. Therefore, it would also be interesting to have a thorough evaluation of the Detection task by investigating users' opinion on each feature of *i*Event during the Detection task. The current findings for the Detection task only investigated the number of features used and data from the interaction logs.

The evaluation of the Tracking tasks was complex since it involved two types of users (postgraduate journalism students and journalists), three components of the interface (Cluster View, Document View and Term View) and two settings (baseline and experimental setup). Thus it affected and limited the evaluation of the Detection task. The availability of the users also contributes to the limitation of the evaluation of the Detection task. This has created another avenue for further research into the usability of the *i*TDT interface for the Detection task.

In addition, other researchers could improve the *i*Event by making it an online TDT system with a variety of current news resources. At present, *i*Event uses a collection of 1998 news stories and resources from CNN. Thus news articles from the New York Times (NYT), CNN or Associated Press (AP) might help with future evaluations.

8.3.3 The Importance of Named Entity Recognition (NER) in *i*TDT

The design of an *i*TDT interface should have an option to display and highlight named entities. This is useful for features such as the cluster labelling and top terms used in Cluster View; especially when users are performing the Tracking task. Users are provided with the two types of information and they can choose to display either keywords or keywords with named entities being highlighted. A combination of this information is interesting since named entities are high quality pieces of information and keywords are descriptive. Rather than providing users with the *Who*, *Where* and *When* it would be interesting to provide the *What* in one setting and investigate its effects.

Applying weighting on named entities in document representation could also improve the system performance. Kumaran and Allan (2004) observed that certain categories of news are better tackled using solely named entities such as Elections, Accidents, Violence and War, New Laws, Sports News and Political and Diplomatic Meetings. For example, the names of election candidates (Person name/*Who*) are very important for stories belonging to the election class. The locations (Location name/*Where*) where accidents happened are also important for stories of within this class. This is the reason why some users found the cluster labelling approach used in Cluster View useful but not effective. In theory, users found that labelling the cluster with the most frequent named entities was useful but not practical when they performed the task. For example Cluster 34 was labelled with the most frequent named entities; *mitch*, *central florida* and *today*, where *today* was not significant enough to tell the user about a topic.

8.4 Chapter Summary

This chapter has concluded the thesis by summarising the contributions made and proposing opportunities to further the research. The findings of this work have fundamental implications for the design of *i*TDT interface and its evaluation. The set of guidelines reported in this thesis is useful for future *i*TDT interface design and the use of

NER in *i*TDT enhanced the effectiveness of users' performance in performing the TDT tasks. Therefore, the contributions made in this work will have benefits for the *i*TDT research community.

Bibliography

1. Allan, J. Introduction to topic detection and tracking. In James Allan, editor, *Topic Detection and Tracking - Event-based Information Organization*, pp. 1-16. Kluwer Academic Publisher, (2002).
2. Allan, J. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norwell, MA. (2002).
3. Allan, J., Lavrenko, V, and Swan, R. Explorations within topic tracking and detection. In James Allan, editor, *Topic Detection and Tracking - Event-based Information Organization*, pp. 197-224. Kluwer Academic Publisher, (2002).
4. Allan, J., Carbonell, J., Doddington, G. R, Yamron, J., and Yang, Y. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218 (1998).
5. Allan, J., Gupta, R. and Khandelal, V. Temporal summaries of news topics. In *Proceedings of the 24rd annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States, pp. 10-18, (2001).
6. Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., and Amstutz, P. Taking Topic Detection From Evaluation to Practice. In *Proceedings of the 38th Annual Hawaii international Conference on System Sciences (HICSS'05) - Track 4 - Volume 04* (January 03 - 06, 2005), IEEE Computer Society, Washington DC, pp. 101a, (2005).

7. Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R. and Caputo, D. Topic-based novelty detection. *Summer workshop at CLSP (Center for Language and Speech Processing) final report*. Baltimore, MD. August. (1999).
8. Allan, J., Lavrenko, V, Malin, D. and Swan, R. Detections, bounds, and timelines: Umass and TDT-3. In *Proceedings of Topic Detection and Tracking (TDT-3)*, pp. 167-174, (2000).
9. Allan, J., Papka, R., and Lavrenko, V. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, ACM Press, pp. 37-45, (1998).
10. Allan, J. Detection as Multi-Topic Tracking. *Information Retrieval*, Kluwer Academic Publisher, 5(2-3): 139-157, (2002).
11. Babbie, E.R. *The Basics of Social Research*. Belmont, CA: Thomson Wadsworth. (2005).
12. Bederson, B. B. Photomesa: a zoomable image browser using quantum treemaps and bubblemaps. In J. Marks and E. Mynatt, editors, *UIST '01: Proceedings of the 14th annual ACM symposium on User interface software and technology*, New York, NY, USA, ACM Press. ISBN 1-58113-438-X, pp. 71–80, (2001).
13. Bederson, B.B., Hollan, J.D., Perlin, K., Meyer, J., Bacon, D., and Furnas, G. *Padd++: A zoomablegraphical sketchpad for exploring alternate interface physics*. *Journal of Visual languages and Computing*, Academic Press, 7(1): 3-31.(1996).
14. Borlund, P. and Ingwersen, P. *The development of a method for the evaluation of interactive information retrieval systems*. *Journal of Documentation*, MCB UP Ltd, 53(3): 225–250. (1997).
15. Borlund, P. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, MCB UP Ltd, 56 (1): 71-90 (2000).
16. Borlund, P. *Evaluation of interactive information retrieval systems*. PhD thesis, Akademi University. (2000).
17. Booch, G. *Object-oriented Analysis and Design with Applications*. Addison-Wesley, 1991.
18. Bouquet, P., Stoermer, H., Niederee, C. and Mana, A. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE*

International Conference on Semantic Computing, ICSC 2008, number CSS-ICSC 2008-4-28-25, IEEE Computer Society, pp. 554-561, (2008).

19. Brants, T. and Chen, F. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 330-337, (2003).
20. Carman, M.J, Crestani, F. Towards personalized distributed information retrieval. *Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2008)*, Singapore, pp. 719-720 (2008).
21. Chen, H and Ku, L.W. A NLP & IR approach to topic detection. In James Allan, editor, *Topic Detection and Tracking - Event-based Information Organization*,. Kluwer Academic Publisher, pp. 243-264, (2002).
22. Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K. and Liberman. M. Corpora for topic detection and tracking. In Allan, J. (ed), *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norwell, MA, pp. 33-66, (2002).
23. Crestani, F., and Ruthven, I. (Eds.). Context: Nature, impact, and role. In *Proceedings of 5th International conference on conceptions of library and information sciences, CoLIS5 2005*. Lecture Notes in Computer Science. Glasgow, UK, vol. 3507. (2005).
24. Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, (2002).
25. Dijk, V. *Discourse Analysis: Its Development and Application to the Structure of News*. The Journal of Communication, 33(2), pp. 20-43 (1983).
26. Doyle, L. *Information Retrieval and Processing*. Wiley, NY. (1975).
27. Eichmann, D and Srinivasan, P. A cluster-based approach to broadcast news. In James Allan, editor, *Topic Detection and Tracking - Event-based Information Organization*, Kluwer Academic Publisher, pp. 149-174, (2002).
28. Eick, S.G., and Wills, G.J. *High Interaction Graphics*. European Journal of Operations Research. 81(3):445-459. (1995)
29. Fiscus, J. G. and Doddington, G. R. Topic detection and tracking evaluation overview. In *Topic Detection and Tracking: Event-Based information Organization*,

- J. Allan, Ed. The Kluwer International Series on Information Retrieval. Kluwer Academic Publishers, Norwell, MA, pp. 17-31 (2002)
30. Franz, M., Todd, W. J., McCarley, S. and Zhu, W.J. Unsupervised and supervised clustering for topic tracking. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001)*, ACM Press, New Orleans, Louisiana, United States, pp. 310-317, (2001).
 31. Gerald Kowalski, *Information Retrieval Systems – Theory and Implementation*, Kluwer Academic Publishers, (1997).
 32. Grishman, R. and Sundheim, B. Message Understanding Conference 6. A Brief History. *Proceedings of the 16th conference on Computational linguistics (COLING-96)*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 466-471, (1996).
 33. Hartigan, J. A. *Clustering Algorithms*. John Wiley & Sons Inc. (1975).
 34. Hearst, M. A. Tilebars: Visualization of term distribution information in full text information access. In Irvin R. Katz, Robert L. Mack, L. M. M. B. R. J. N., editor, *Proceedings of the ACM CHI 95 Human Factors in Computing Systems Conference*, ACM Press/Addison-Wesley Publishing Co., Denver, Colorado, United States, pp. 59–66. (1995).
 35. Hearst, M. *Search User Interfaces*, Cambridge University Press, September (2009).
 36. Ingwersen, P. and Järvelin, K. *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of Series: The Information Retrieval Series. Springer, ISBN: 1-4020-3850-X. (2005).
 37. Ingwersen, P., & Järvelin, K. *Second Workshop on Information Retrieval in Context (IRiX)*. ACM SIGIR 2005, Salvador, Brazil. (2005).
 38. Jones, G. J. F. and Gabb, S. M. A visualisation tool for topic tracking analysis and development. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM, New York, NY, 389-390 (2002).
 39. Kovach, B. and Rosenstiel, T. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. Crown Publishers, New York (2001).
 40. Kumaran, G. and Allan, J. Text Classification and Named Entities for New Event Detection. *Proceedings of the 27th Annual International ACM SIGIR Conference*, New York, NY, USA. ACM Press. pp. 297–304. (2004).

41. Kuo, Z., Zi, L.J. and Gang, W. New Event Detection Based on Indexing-tree and Named entities. *Proceedings of SIGIR '07*, ACM, Amsterdam, The Netherlands, pp. 215-222. (2007).
42. Leuski, A., and Allan, J. Filtering: Improving realism of topic tracking evaluation. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 89-96, (2002).
43. Leuski, A., and Allan, J. Interactive cluster visualization for information retrieval. In *Proceedings of European Conference on Digital Libraries (ECDL '98)*, Heraklion, Crete, Greece, Lecture Notes in Computer Science, Springer, pp. 535-554 (1998).
44. Leuski, A., and Allan, J. Lighthouse: Showing the Way to Relevant Information. In *Proceedings of the IEEE Symposium on information Visualization 2000*. INFOVIS. IEEE Computer Society, Washington, pp. 125-129 (2000).
45. Lewis D. and Gale, W. A Sequential Algorithm for Training Text Classifiers, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc. Dublin, Ireland, pp. 3-13, (1994).
46. Li, X. and Croft, W.B. Novelty Detection Based on Sentence Level Information Patterns. *Proceedings of ACM Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany, 31st October - 5th November, ACM, pp. 744-751 (2005).
47. Li, X. Sentence Level Information Patterns for Novelty Detection. Ph.D thesis. University of Massachusetts at Amherst. (2006).
48. Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. Applying Semantic Classes in Event Detection and Tracking. *Proceedings of International Conference on Natural Language Processing (ICON 2002)*, Mumbai, India, pp. 175-183. (2002).
49. Makkonen, J., Ahonen-Myka, H. and Salmenkivi, M. Simple Semantics in Topic Detection and Tracking. *Information Retrieval*, Springer, 7(3-4): 347-368. (2004).
50. Nallapati, R., Feng, A., Peng, F., and Allan, J. Event threading within news topics. In *Proceedings of the Thirteenth ACM international Conference on information and Knowledge Management (Washington, D.C., USA, November 08 - 13, 2004)*. CIKM '04. ACM, New York, NY, pp. 446-453 (2004).
51. Osgood, C.E., Suci, G., & Tannenbaum, P. *The measurement of meaning*. Urbana, IL: University of Illinois Press. (1957).

52. Otterbacher, J. Erkan, G. and Radev, D.R. Using Random Walks for Question focused Sentence Retrieval. *Proceedings of Human Language Technology / Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, Vancouver, Canada, 6-8 October, pp. 915–922. ACM. (2005).
53. Papka, R and Allan, J. On-line new event detection using single-pass clustering. *Technical Report IR-123*, Department of Computer Science, University of Massachusetts. (1998).
54. Pons-Porrata, A. Berlanga-Llavori, R. Ruiz-Shulcloper, J. Perez-Martinez, J. M. JERARTOP: A New Topic Detection System. In *Proceeding of Progress in Pattern Recognition, Image Analysis and Applications*. Mexico. Lecture Notes in Computer Science. Volume 3287/2004. pp. 446-453 (2004).
55. Ruthven, I. Recasting the context in information retrieval. *Second International Workshop on Information retrieval, Logic and Uncertainty*. University of Glasgow, pp. 68-72. (1996).
56. Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N. J., Tombros, A., and Vakkari, P. In *Proceedings of the 1st international conference on Interaction in context*. ACM Press, vol. 176. (2006).
57. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, Pergamon Press, Inc. 24(5): 513–523. (1988).
58. Sekine, S., and Isahara, H. IREX: IR and IE Evaluation project in Japanese, *Proceedings of International Conference on Language Resources & Evaluation (LREC 2000)*, Athens, Greece, (2000).
59. Sekine. S, Sudo, K., and Nobata, C. Extended Named entities Hierarchy. *Proceedings of International Conference on Language Resources & Evaluation (LREC 2000)*, Athens, Greece, (2002).
60. Shneiderman, B. and Aris, A., Network Visualization by Semantic Substrates, *Proceedings of IEEE Visualization/Information Visualization. IEEE Transactions on Visualization and Computer Graphics* 12(5): 733-740, (2006).
61. Sheiderman, B and Plaisant, C. The future of graphic user interfaces: Personal rolemanagers. In *People and Computers IX, British Computer Society HCI'94*, Cambridge University Press, pp. 3–8, (1994).
62. Shneiderman, B. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley, Reading, M.A, (1997).

63. Spärck-Jones, K. *Information retrieval experiment*. London Butterworths (1981).
64. Spärck-Jones, K. and Willet, P. *Readings in Information Retrieval*. Morgan Kaufmann Publishers. (1997).
65. Stovall, J.G. *Journalism: Who, What, When, Where, Why and How*. Allyn & Bacon. Boston, MA. (2004).
66. Swan, R. and Allan, J. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM, New York, NY, pp. 49-56 (2000).
67. Teevan, J. *The Re-Search Engine: Helping People Return to Information in Dynamic Information Environments*. Ph.D. Massachusetts Institute of Technology. (2007).
68. Vallet, D., Fernández, M., Castells, P., Mylonas, Ph., and Avrithis, Y. Personalized Information Retrieval in Context. 21st National Conference on Artificial Intelligence - 3rd International Workshop on Modeling and Retrieval of Context, Boston, USA, (2006).
69. Van Rijsbergen, C. J. *Information retrieval*, 2ed., Butterworths, London, (1979).
70. Voorhees, E.M. and Harman, D. Overview of the Eighth Text REtrieval Conference (TREC-8). *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, Gaithersberg, Maryland, 17-19 November, NIST Special Publication, pp. 1–24, (2000).
71. White, R. W., Jose, J. M. and Ruthven, I. An implicit feedback approach for interactive information retrieval. *Information Processing and Management*. 42 (1). Pergamon Press Inc. pp. 166-190. (2006).
72. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B.T. and Liu, X. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval*, IEEE Educational Activities Department, 14(4):32-43. (1999).
73. Yang, Y., Zhang, J., Carbonell, J. and Jin, C. Topic conditioned Novelty Detection. *Proceedings of the 8th ACM SIGKDD International Conference*, ACM Press. pp. 688–693. (2002).

Published Work

1. Mohd, M., Crestani, F. and Ruthven, I. Design of an Interface for Interactive Topic Detection and Tracking. 8th International Conference on Flexible Query Answering Systems (FQAS 2009). pp. 227–238. Roskilde, Denmark (2009).
2. Mohd, M., Crestani, F. and Ruthven, I. A comparison of Named Entity Patterns from a User Analysis and a System Analysis. 30th European Conference on Information Retrieval (ECIR 2008). pp. 679-683 (2008).
3. Mohd, M. Named Entity Patterns across News Domains. BCS IRSG Symposium: Future Directions in Information Access 2007. Glasgow, Scotland. (2007).

Appendix A

Resources relating to the work described in Chapter 3

- A.1 Information sheet, consent form and receipt of payment
- A.2 Entry questionnaire
- A.3 Tracking task questionnaire
- A.4 Detection task questionnaire
- A.5 *i*Event post evaluation survey

Appendix A.1

Information sheet, consent form and receipt of payment

Department: *Computer and Information Sciences*
Subject Identification Number for this study:



INFORMATION SHEET

Title of Project: **Event Interface Evaluation**
Name of researcher: Masnizah Mohd

You are being invited to take part in a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully. Ask me if there is anything that is not clear or if you would like more information.

The aim of this experiment is to investigate the effectiveness of Event (News Monitoring System) interface in tracking and detecting a topic. We cannot determine the value of the system interface unless we ask those people who are likely to be using them, which is why we need to run experiments like these. Please remember that it is the interface, not you, that are being evaluated. You were chosen, along with 20 others, because you are the postgraduate journalism student or a journalist.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. If you decide not to take part you are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed. A decision not to participate will not affect your grades in any way. The experiment will take about 3 hours excluding a short training session and you will receive a reward of £20 upon completion. You will be given a chance to learn how to use the interface before we begin. At this time you will also be asked to complete an introductory questionnaire.

There are 2 tasks in this experiment which are TOPIC TRACKING and TOPIC DETECTION tasks. In TOPIC TRACKING task, you will perform 8 topic tracking tasks in two sessions and complete a questionnaire about using the interface. The two sessions should not take you more than 2 hours and you have been asked to spend 15 minutes for each TOPIC TRACKING task. The questionnaires will ask how you felt during each search.

In TOPIC DETECTION tasks, you will perform 4 topic detection tasks in two sessions and complete a questionnaire about using the interface. The two sessions should not take you more than 1 hour and you have been asked to spend 10 minutes for each TOPIC DETECTION task.

All of your interaction (e.g., mouse clicks, key presses) will also be logged. You are encouraged to comment on each interface as you use it. Please ask questions if you need to and please let me know when you are finished each task. You will be asked some questions about the tasks and systems at the end of the experiment.

All information which is collected about you during the course of this research will be kept strictly confidential. You will be identified by an ID number and all information about you will be removed so that you cannot be recognised from it. Data will be stored only for analysis, and then destroyed. Data collected during the study will be statistically analysed and presented in various forms, including quotations. The results of this study will be used for my Ph.D. research and the data may be published in a thesis, research papers or presentation. The results are likely to be published in late 2009. You will not be identified in any report or publication that arises from this work. This study too has obtained the ethical consent. For further information about this experiment please contact:

Masnizah Mohd (e.mail: Masnizah.Mohd@cis.strath.ac.uk or tel: 0141 548 3583)

1/6/09

Department: *Computer and Information Sciences*
Subject Identification Number for this study:



CONSENT FORM

Title of Project: **Event Interface Evaluation**
Name of researcher: Masnizah Mohd

Please initial box

1. I confirm I have read and understand the information sheet dated (01/06/09) for the above study and have had the opportunity to ask questions.

2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected.

3. I agree to take part in the above study.

Name of subject	Date	Signature
-----------------	------	-----------

Researcher	Date	Signature
------------	------	-----------

1 for subject; 1 for researcher

Department: *Computer and Information Sciences*
Subject Identification Number for this study:



RECEIPT OF PAYMENT

Title of Project: **Event Interface Evaluation**
Name of researcher: Masnizah Mohd

I confirm receipt of £20 paid for my participation in the above experiment.

Name of subject	Date	Signature
-----------------	------	-----------

Researcher	Date	Signature
------------	------	-----------

Appendix A.2

Entry questionnaire

ENTRY QUESTIONNAIRE

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment.



ID:

Please place a TICK in the square that best matches your opinion.

PERSONAL INFORMATION

1. Please indicate your AGE:

Below 20	<input type="checkbox"/>	1
20 – 30 years	<input type="checkbox"/>	2
30 - 40 years	<input type="checkbox"/>	3
Above 40	<input type="checkbox"/>	4

2. Please indicate your GENDER:

Male	<input type="checkbox"/>	1
Female	<input type="checkbox"/>	2

3. Please indicate your EDUCATION LEVEL :

School (Standard / O grade)	<input type="checkbox"/>	1
School (Standard / A grade)	<input type="checkbox"/>	2
College (HNC)	<input type="checkbox"/>	3
College (HND)	<input type="checkbox"/>	4
University (Undergraduate Degree)	<input type="checkbox"/>	5
University (Postgraduate Degree)	<input type="checkbox"/>	6
Other	<input type="checkbox"/>	7

WORK EXPERIENCE

4. Do you have an experience working as a JOURNALIST before? :

Yes 1
 No 2

If YES,

Less than 2 years	<input type="checkbox"/>	1
2 to 5 years	<input type="checkbox"/>	2
6 to 10 years	<input type="checkbox"/>	3
Other (please specify)	<input type="checkbox"/>	4

4.1 Please indicate your type of journalist:

Feature/ Magazine writer	<input type="checkbox"/>	1
Daily news reporter	<input type="checkbox"/>	2
Other (please specify)	<input type="checkbox"/>	3

NEWS SEARCHING EXPERIENCE

5. Please indicate which news network, tools or search engines you use (mark as MANY as apply):

CNN	<input type="checkbox"/>	1
BBC News	<input type="checkbox"/>	2
Al Jazeera	<input type="checkbox"/>	3
SkyNews	<input type="checkbox"/>	4
JournalismNet	<input type="checkbox"/>	5
Journalism.co.uk	<input type="checkbox"/>	6
NewsTrust.net	<input type="checkbox"/>	7
RocketNews	<input type="checkbox"/>	8
Google	<input type="checkbox"/>	9
Yahoo	<input type="checkbox"/>	10
Altavista	<input type="checkbox"/>	11
AlltheWeb	<input type="checkbox"/>	12
Others (please specify)	<input type="checkbox"/>	13

USER TASKS

To evaluate the *Æ*Event interface, we now ask you to answer some questions about them. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers.

In this task, you are to imagine that you work for a newswire agency that is responsible for reporting news. Your particular role is to monitor the news media for information; track and detect about such events or topics. Part of the task is to report a story and to identify the topic.

You will be using *Æ*Event, a news monitoring system with Topic Detection and Tracking (TDT) capabilities which aims to cluster news stories into the same group of events or topics by visualizing the clusters.

Please use only the information expressed in this document and the information given to you from the *Æ*Event interface output.

You are advised to use the features in each component of the *Æ*Event interface.

You will be given a set of news articles related to the event, published by CNN news sources at different points in time and the summary of each topic.

Please remember that we are evaluating the system interface you have just used and not you.

Appendix A.3

Tracking task questionnaire

TOPIC 1

OPRAH LAWSUIT

Place a TICK in the square that best matches your opinion. Please answer all questions.

TITLE: Oprah Lawsuit

TOPIC SUMMARY

It is January 1998. During a recent television programme about Mad-Cow disease Oprah Winfrey announced 'I will never eat beef again' due to concerns about the safety of American beef production. As a result the sales of US beef dropped and the Cattle Industry sued Oprah for damaging their industry. The trial happened in Amarillo, Texas, and has now finished with a victory for Oprah.


You have been asked to write an article on the outcome of this legal action against Oprah Winfrey.

Topics likely to be important are Oprah's legal teams preparations; the testimony from either side; and reactions to the verdict.

SECTION 1.1: TOPIC FAMILIARITY


Please answer this question [Section 1] before you start with the system.

1. How familiar are you with this topic?

Not at all  Extremely familiar

1 2 3 4 5

2. How interested are you with this topic?

Not at all  Very interested

1 2 3 4 5

SECTION 1.2: TOPIC TRACKING

You may now start using the system.

3. If you are given a task to write an article on the outcome of this legal action against Oprah Winfrey, please draft the important facts or points in a point form (use over leaf if necessary).

4. Please list any useful cluster/s for tracking information on this topic.


5. You would like to track and keep new information on this topic. In order to track new information, you need to create a profile of useful keywords which will highlight new documents. Please create such a profile below. You may list up to 10 keywords.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____

6. Using this interface to track the topic was GENERALLY:


	1	2	3	4	5	
Difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Easy
Stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Relaxing
Complex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Simple
Frustrating	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Satisfying
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

7. I believe I have gathered enough information on this topic.

Disagree  Agree


<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

8. After getting the information, how familiar are you with this topic NOW?

Not at all  Extremely familiar

1 2 3 4 5

9. After getting the information, how interested are you with this topic NOW?

Not at all  Very interested

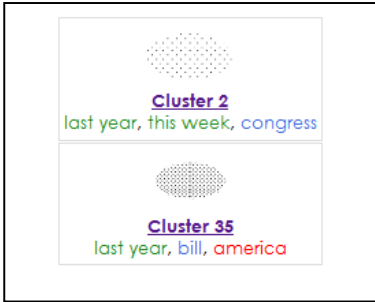
1 2 3 4 5

10. Do you have any further comments?

SECTION 2: FEATURES

In this section we ask you about the interface you have just used.

11. For this topic, the keyword approach used in Cluster Labelling (Cluster View) was:



	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

12. For this topic, the keyword approach used in Top Terms (Cluster View) was:

Top Terms for Cluster 1:

1. u.s. 2. birmingham 3. new york 4. kaczyński 5. today 6. alabama 7. cnn 8. fbi 9. police 10. atlanta

	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

13. For this topic, the cluster visualisation used in Cluster View was:



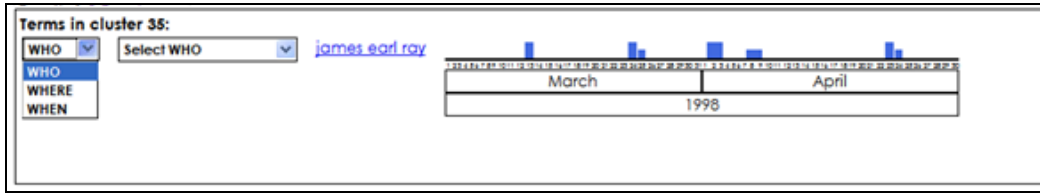
	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

14. For this topic, the cluster description button used in Cluster View was:

Cluster	Description
	Cluster with large size and high density. This cluster contains high number of documents with short period of time.
	Cluster with large size and medium density. This cluster contains high number of documents with medium period of time.
	Cluster with large size and low density. This cluster contains high number of documents with long period of time.
	Cluster with medium size and high density. This cluster contains medium number of documents with short period of time.
	Cluster with medium size and medium density. This cluster contains medium number of documents with medium period of time.
	Cluster with medium size and low density. This cluster contains medium number of documents with long period of time.
	Cluster with small size and high density. This cluster contains small number of documents with short period of time.
	Cluster with small size and medium density. This cluster contains small number of documents with medium period of time.
	Cluster with small size and low density. This cluster contains small number of documents with long period of time.

	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

15. For this topic, the keyword approach used in Term View was:



Not useful	<input type="text" value="1"/>	<input type="text" value="2"/>	<input type="text" value="3"/>	<input type="text" value="4"/>	<input type="text" value="5"/>	Useful
Ineffective	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Effective
Unhelpful	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Helpful
Boring	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Interesting

16. For this topic, the document content used in Document View was:

Documents in cluster 1:

1. [CNN19980104.1130.0453](#)
2. [CNN19980105.1130.0033](#)
3. [CNN19980105.1130.0898](#)
4. [CNN19980105.1600.0034](#)
5. [CNN19980106.0100.0026](#)
6. [CNN19980108.1130.0086](#)
7. [CNN19980108.1130.0965](#)
8. [CNN19980109.1130.0082](#)
9. [CNN19980112.1130.0246](#)
10. [CNN19980129.1130.0074](#)

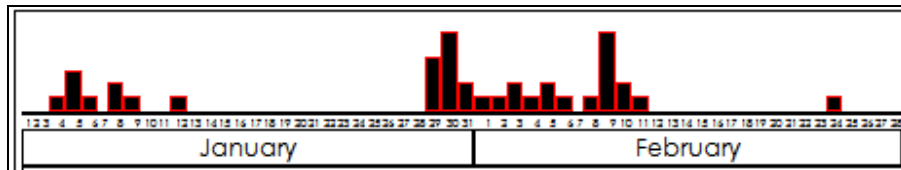
Result: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

12 January 1998 | CNN19980112.1130.0246

They begin testing defendant [Ted Kaczynski](#) to determine if he's com rejected another offer from [Kaczynski's](#) lawyers to plead guilty to a reconsider its decision to seek the death penalty. [Rusty Dornin](#) has n prosecutors might reconsider a plea bargain that would sentence th government would look much more favorably on that offer. As neg circus, and questions about [Kaczynski's](#) mental state, might make sympathize with [Ted Kaczynski](#) and the mother and the brother and have accepted the offer from [Kaczynski](#) to plead guilty for a life set plea bargain unless [Kaczynski](#) is proven competent. [This week](#) the mental exam. His examiner, [Dr. Sally Johnson](#), seen in this [1981](#) cour assailant [John Hinckley](#) and found him to be sane at the time he sh competency testing has different standards than those of insanity. If mental disease that would render him incapable of forming the me himself and present himself at trial. The exam comes only days after judge [Garland Surrell](#) seems determined to keep the case from fur stand by -- [January](#) 22nd. [Rusty Dornin](#) reporting.

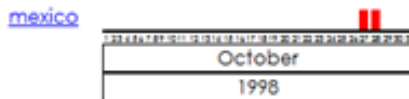
Not useful	<input type="text" value="1"/>	<input type="text" value="2"/>	<input type="text" value="3"/>	<input type="text" value="4"/>	<input type="text" value="5"/>	Useful
Ineffective	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Effective
Unhelpful	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Helpful
Boring	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Interesting

17. For this topic, the histogram and the timeline used in Document View were:



	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

18. For this topic, the histogram and the timeline used in Term View were:



	1	2	3	4	5	
Not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Useful
Ineffective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Effective
Unhelpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Helpful
Boring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Interesting

Appendix A.4

Detection task questionnaire

SECTION 3: TOPIC DETECTION

1. Please indicate the topic that Cluster 52 is dealing with. If you think there is more than one topic in this cluster, please rank maximum 3 topics in the list below:

	Topic Name	Rank
1.	Yeltsin's Illness	
2.	Gingrich Resigns	
3.	Diane Zamora	
4.	Clinton-Jiang Debate	
5.	Sgt Gene McKinney	
6.	Osama bin Laden Indictment	
7.	Bombing AL Clinic	
8.	Joe DiMaggio Illness	
9.	Hurricane Mitch	
10.	Space Station Module Zaria Launched	
11.	Microsoft Anti-Trust Case	
12.	NBA finals	
13.	Grossberg baby murder	
14.	Monica Lewinsky Case	
15.	American Embassy Bombing Trial	
16.	Rats in Space!	
17.	Yankees vs. Padres in World Series	
18.	House Speaker-Elect Livingston Resigns	
19.	US Mid-term Elections	
20.	Clinton's Gaza Trip	

2. How easily could you identify these topics?

Difficult 1 2 3 4 5 Easy

3. Please indicate which features in this interface help you in identifying the topic (mark as MANY as apply):

Keyword approach used in Cluster Labelling (Cluster View)



Cluster 2
last year, this week, congress

Cluster 35
last year, bill, america

Keyword approach used in Top Terms (Cluster View)

Top Terms for Cluster 1:
1. u.s. 2. birmingham 3. new york 4. kaczynski 5. today 6. alabama 7. cnn 8. fbi 9. police 10. atlanta

Cluster visualisation used in Cluster View



Cluster 37 Cluster 38 Cluster 39 Cluster 40

Keyword approach used in View



Terms in cluster 35:
WHO: Select WHO james earl ray
WHERE: [empty]
WHEN: [empty]

March April 1998

Document content used in Document View



Documents in cluster 1:
12 January 1998 | CNN19980112.1130.0246

1. CNN19980106.1130.0452 They begin testing defendant **Ray** to determine if he's co-
2. CNN19980106.1130.0528 rejected another offer from **Attorney** to plead guilty to a
3. CNN19980106.1130.0528 reconsider its decision to seek the death penalty. **Ray** has
4. CNN19980106.1600.0504 prosecutors might reconsider a plea bargain that would sentence
5. CNN19980106.0500.0026 government would look much more favorably on that offer. As the
6. CNN19980106.1130.0056 circus, and questions about **Ray's** mental state, might make
7. CNN19980106.1130.0745 sympathize with **Ray** and the mother and the brother on
8. CNN19980106.1130.0026 have accepted the offer from **Attorney** to plead guilty for a life as
9. CNN19980112.1130.0246 plea bargain unless **Ray** is proven competent. **Ray** was
10. CNN19980106.1130.0026 the mental exam. His examiner, **Dr. [Name]**, seen in the **trial** court
Result: 1 2 3 4 5 6 7 8 9 10

12 January 1998 | CNN19980112.1130.0246
They begin testing defendant **Ray** to determine if he's co-
rejected another offer from **Attorney** to plead guilty to a
reconsider its decision to seek the death penalty. **Ray** has
prosecutors might reconsider a plea bargain that would sentence
government would look much more favorably on that offer. As the
circus, and questions about **Ray's** mental state, might make
sympathize with **Ray** and the mother and the brother on
have accepted the offer from **Attorney** to plead guilty for a life as
plea bargain unless **Ray** is proven competent. **Ray** was
the mental exam. His examiner, **Dr. [Name]**, seen in the **trial** court
evaluated **Ray's** and found him to be sane at the time he
competency testing has different standards than those of insanity.
mental disease that would render him incapable of forming the
himself and present himself at trial. The exam comes only days after
judge **Richard [Name]** seems determined to keep the case from fur
stand by **Attorney** and **Attorney** reporting.

Histogram and the timeline used in Document View



January February

Histogram and the timeline used in List View



mexico
October 1998

Others (please specify)

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Appendix A.5

*i*Event post evaluation survey

IEVENT POST EVALUATION SURVEY

These questions intend to investigate user experiences after using both interfaces in performing the tasks by an interview.

1. How did you feel about each interface you used?

2. Did you notice any difference using both interfaces?

3. In TOPIC TRACKING, please indicate which interface helps you to perform the following task:

a. Tracking the topics.

SETUP 1	SETUP 2	NOT SURE
---------	---------	----------

Why:

Reporting a story.

SETUP 1	SETUP 2	NOT SURE
---------	---------	----------

Why:

b. Creating a profile for a topic.

SETUP 1	SETUP 2	NOT SURE
---------	---------	----------

Why:

4. In TOPIC DETECTION, please indicate which interface helps you to perform the following task:

a. Detecting the topics.

SETUP 1	SETUP 2	NOT SURE
---------	---------	----------

Why: