JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Research Paper
J. of infosci. theory and practice 2(4): 31-48, 2014
http://dx.doi.org/10.1633/JISTaP.2014.2.4.3

Investigating the Combination of Bag of Words and Named Entities Approach in Tracking and Detection Tasks among Journalists

Masnizah Mohd *
Japan Advanced Institute of Science and Technology (JAIST), Japan
E-mail: masnizah@jaist.ac.jp

Omar Mabrook A. Bashaddadh
Universiti Kebangsaan Malaysia, Malaysia
E-mail: aboshdad09@hotmail.com

ABSTRACT

The proliferation of many interactive Topic Detection and Tracking (*i*TDT) systems has motivated researchers to design systems that can track and detect news better. *i*TDT focuses on user interaction, user evaluation, and user interfaces. Recently, increasing effort has been devoted to user interfaces to improve TDT systems by investigating not just the user interaction aspect but also user and task oriented evaluation. This study investigates the combination of the bag of words and named entities approaches implemented in the *i*TDT interface, called Interactive Event Tracking (*i*Event), including what TDT tasks these approaches facilitate. *i*Event is composed of three components, which are Cluster View (CV), Document View (DV), and Term View (TV). User experiments have been carried out amongst journalists to compare three settings of *i*Event: Setup 1 and Setup 2 (baseline setups), and Setup 3 (experimental setup). Setup 1 used bag of words and Setup 2 used named entities, while Setup 3 used a combination of bag of words and named entities. Journalists were asked to perform TDT tasks: Tracking and Detection. Findings revealed that the combination of bag of words and named entities approaches generally facilitated the journalists to perform well in the TDT tasks. This study has confirmed that the combination approach in *i*TDT is useful and enhanced the effectiveness of users' performance in performing the TDT tasks. It gives suggestions on the features with their approaches which facilitated the journalists in performing the TDT tasks.

Keywords: Bag of Words, Named Entity Recognition, Interactive Topic Detection and Tracking, User, Journalists

Open Access

Accepted date: October 23, 2014
Received date: July 30, 2014

***Corresponding Author:** Masnizah Mohd
Postdoctoral researcher
Japan Advanced Institute of Science and Technology (JAIST), Japan
E-mail: masnizah@jaist.ac.jp

1. INTRODUCTION

A vast amount of information arrives every day through a variety of news media such as newswires, TV, newspapers, radio, and website sources. In addition, the significant growth and dynamic environment of digital information creates challenges in information retrieval (IR) technology. The amount of constant information which readers can effectively use is still limited compared to the continuous growth of rapidly changing online information. Thus it has become increasingly important to provide and enhance methods to find and present textual information effectively and efficiently. Technologies have been proposed to solve the information overload issues, including information customization, search engines, and information agency. Therefore research on Topic Detection and Tracking (TDT) serves as a core technology for a system that monitors news broadcasts. It helps to alert information professionals such as journalists to interesting and new events happening in the world. Journalists are keen to track and, in particular, to know the latest news about a story from a large amount of information that arrives daily.

Most TDT research has concentrated primarily on the design and evaluation of algorithms to implement TDT tasks such as stream segmentation, story detection, link detection, and story tracking. Bag of words (BOW) and named entities (NE) approaches are widely implemented either in document representation or in user interface design of TDT systems (Kumaran & Allan, 2004). Meanwhile Interactive Topic Detection and Tracking (*i*TDT) refers to the TDT works which focus on user interaction, user evaluation, and user interfaces aspects (Mohd et al., 2012).

In this research we did not discuss the interface design or features but focused on the approaches used: either BOW or NE or their combination. The objectives of this research are: (i) to implement the combined approach of BOW and NE in *i*TDT interface and, (ii) to investigate the usability of this approach on the *i*TDT interface in assisting journalists to perform the TDT tasks. This is important to answer questions, e.g., which approach aims to facilitate journalists to perform TDT tasks? Will the use of a combined approach of BOW and NE improve *i*TDT? What are the TDT tasks facilitated through the use of this combined approach?

2. RELATED WORK

The focus of this section is to review related works on *i*TDT with a view to identifying best approaches used in the design of *i*TDT interfaces. The previous scholarly studies in TDT tried to build better document models, enhancing and developing similarity metrics or document representations. A large number of research efforts have focused on enhancing and improving document representations by applying Natural Language Processing (NLP) technology such as Named Entity Recognition (NER) (Yang et al., 1999, Makkonen et al., 2004, Kumaran & Allan, 2004, Zhang et al., 2004). Some research tended towards interactive TDT through casting its attention on the graphical user interface (GUI), rather than on laboratory experiments such as TDT Lighthouse (Leuski & Allan, 2000), TimeMine (Swan & Allan, 2000), Topic Tracking Visualization Tool (Jones & Gabb, 2002), Event Organizer (Allan et al., 2005), Stories in Time (Berendt & Subasic, 2009) and Interactive Event Tracking System (*i*Event) interface (Mohd et al., 2012), which are all examples of TDT studies that investigate some approaches to enhancing and improving TDT systems' and users' performance. Based on the works reviewed, none of them investigate the effectiveness of their interfaces using the combination of a bag of words (BOW) and named entities (NE) approach. We review and discuss the approaches used in these related studies and we concentrate in *i*Event as the baseline setup.

2.1. Interactive Event Tracking (*Ievent*) Interface

The Interactive Event Tracking System (*i*Event) interface (Mohd et al., 2012) is a prototype system that includes the design of a novel *i*TDT system interface to facilitate professionals such as journalists and information analysts in performing TDT tasks. This system consists of two settings: Setup 1 is the baseline Setup that uses the BOW approach and Setup 2 is the experimental Setup that uses the NE approach to present the information to the user in a meaningful way using the following components: Cluster View (CV), Document View (DV), and Term View (TV).

*i*Event incorporates a number of components and features into a single interface, such as cluster labelling, top terms, histogram with the timeline, document content, and keyword. Moreover, it implements good approaches, such as BOW and NE. However, these approaches are only implemented separately, each one in a single setup. Clearly the BOW offers a better explanation in the Detection task as it resembles the What. Meanwhile NE provides specific information on the Who, Where, and When in performing the Tracking task (Mohd et al., 2012). The major problem with this kind of application and approach is that neither supports both tasks (Detection and Tracking) to gather information. So this research attempts to solve this problem by adding a new Setup called Setup 3. This new Setup, which will be considered as an experimental Setup which uses a combination approach (BOW+NE), is intended to offer a better explanation and provide information on the *Who, Where, When,* and *What* to the user.

The reviewed works in *i*TDT used different approaches to extract, organize, and represent information from the news sources to the user such as BOW, NE, query, graphical visualization, text summarization, and other approaches. But none of these works supports tracking and detection tasks at the same time. Therefore, the combination of BOW and NE could enhance user performance in TDT tasks.

3. THE USE OF BAG OF WORDS AND NAMED ENTITIES IN INTERACTIVE EVENT TRACKING SYSTEM (*i*EVENT) INTERFACE

There are four features out of seven of *i*Event which distinguish between three approaches: BOW, NE, and a combination of BOW and NE across setups, as apparent in Table 1. The descriptions are:

a. BOW are the keywords aiming to provide the *What* and were implemented in Setup 1. Most frequent keywords were offered in the features of *i*Event.
b. NE are the significant keywords that aim to provide the *Who, Where* and *When* and were implemented in Setup 2. Most frequent named entities were offered in the features of *i*Event.
c. Combination of BOW and NE that aim to provide the *What, Who, Where* and *When* and were implemented in Setup 3. Most frequent keywords and named entities were offered in the features of *i*Event.

The four features of *i*Event that are different between the setups are: 'CV: cluster labelling,' 'CV: top terms,' 'DV: document content,' and 'TV: keyword approach.' Each was using a combination of BOW and NE in Setup 3 (experimental setup) and using BOW alone in Setup 1 (baseline setup) and NE alone in Setup 2 (baseline setup). We extracted the named entities using ANNIE (A Nearly-New Information Extraction system), which is an information extraction component of the General Architecture for Text Engineering (GATE). We used it for its accurate entity, pronoun, and nominal co-references extraction (Cunningham et al., 2002).

Table 1. The Differences of the Three Setups

	Baseline Setup		**Experimental Setup**
	Setup 1	**Setup 2**	**Setup 3**
CLUSTER VIEW (CV)			
cluster labelling	BOW	NE	BOW+NE
top terms	BOW	NE	BOW+NE
cluster visualization	×	×	×
DOCUMENT VIEW (DV)			
histogram with the timeline	×	×	×
document content	BOW	NE	BOW+NE
TERM VIEW (TV)			
keyword approach	BOW	NE	BOW+NE
histogram with the time line	×	×	×

NE=Named Entities; BOW=Bag of words
×=unavailable

3.1. Cluster View (CV)

The Cluster View visualized information related to the size and density of a cluster, and displayed the ten most frequent terms (Top terms) in the cluster. The main difference here is in the Cluster labelling feature (refer to Fig. 1) and top ten terms feature (refer to Fig. 2). Clusters were generated using single pass clustering (Mohd et al., 2012) and were labelled using the three most frequently named entities; for example, Cluster 30 was labelled using keywords in Setup 1, named entities in Setup 2, and a combination of BOW and NE in Setup 3.

When the user clicked on Cluster 30, there was additional information presented on the top ten terms as shown in Fig. 2. The difference in these features between the three setups is that the user was provided with mixed BOW and NE in Setup 3 instead of using BOW only in Setup 1 and NE only in Setup 2.

3.2. Document View (DV)

Document View provided the user with information about the document timeline in a histogram form and the document content in a cluster. There was no difference in timeline across setups where this feature was showing the occurrence and the number of documents (document frequency) for a specific date. This supported the journalist in viewing the information flow in a press article or in analyzing the discourse.

However, as shown in Fig. 3, users could see a small difference in the document content, especially between Setup 3 (experimental Setup) and Setup 2 (baseline Setup) and a significant difference with Setup 1 (baseline Setup). The types of term such as Who, What, Where, and When are highlighted according to the setup.

Cascading Style Sheets (CSS) were used to differentiate three types of named entities with different colors and terms assigned, as shown in Table 2.

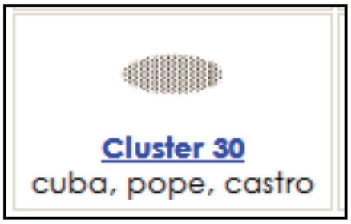
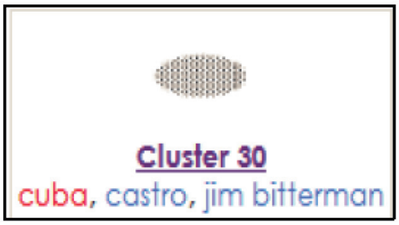
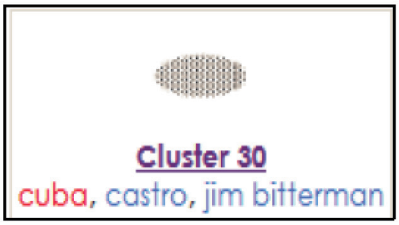
BOW (Setup 1)	Named Entities (Setup 2)	BOW+NE (Setup 3)
 Cluster 30 cuba, pope, castro	 Cluster 30 cuba, castro, jim bitterman	 Cluster 30 cuba, castro, jim bitterman

Fig.1 Cluster labelling of the three approaches

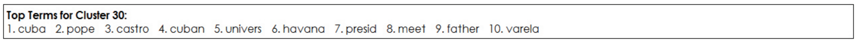
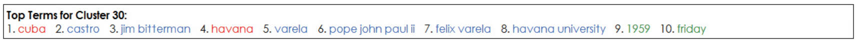
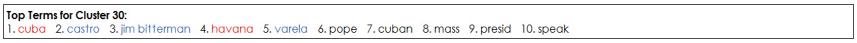
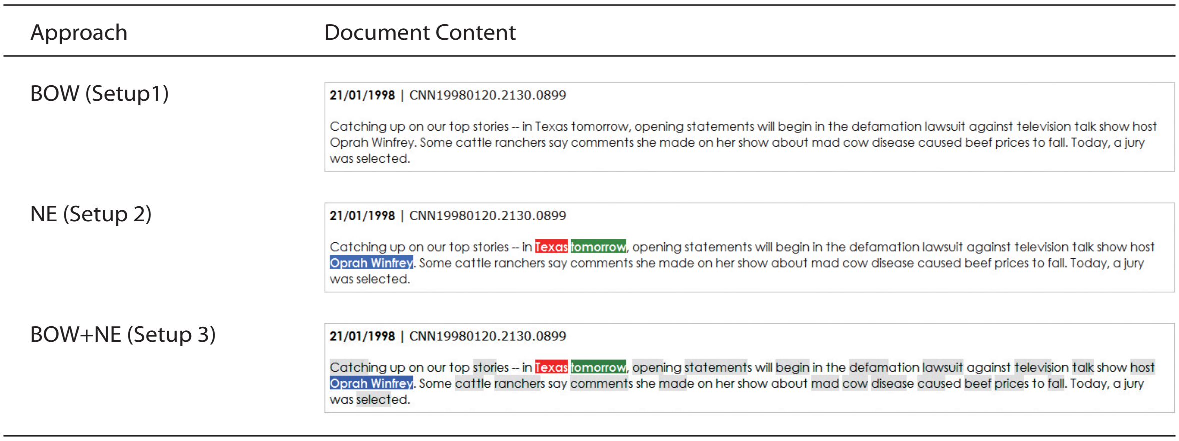
Approach	Top Ten Terms
BOW (Setup1)	 Top Terms for Cluster 30: 1. cuba 2. pope 3. castro 4. cuban 5. univers 6. havana 7. presid 8. meet 9. father 10. varela
NE (Setup 2)	 Top Terms for Cluster 30: 1. cuba 2. castro 3. jim bitterman 4. havana 5. varela 6. pope john paul ii 7. felix varela 8. havana university 9. 1959 10. friday
BOW+NE (Setup 3)	 Top Terms for Cluster 30: 1. cuba 2. castro 3. jim bitterman 4. havana 5. varela 6. pope 7. cuban 8. mass 9. presid 10. speak

Fig. 2 Top ten terms of the three approaches



Approach	Document Content
BOW (Setup1)	21/01/1998 | CNN19980120.2130.0899 Catching up on our top stories -- in Texas tomorrow, opening statements will begin in the defamation lawsuit against television talk show host Oprah Winfrey. Some cattle ranchers say comments she made on her show about mad cow disease caused beef prices to fall. Today, a jury was selected.
NE (Setup 2)	21/01/1998 | CNN19980120.2130.0899 Catching up on our top stories -- in Texas tomorrow, opening statements will begin in the defamation lawsuit against television talk show host Oprah Winfrey. Some cattle ranchers say comments she made on her show about mad cow disease caused beef prices to fall. Today, a jury was selected.
BOW+NE (Setup 3)	21/01/1998 | CNN19980120.2130.0899 Catching up on our top stories -- in Texas tomorrow, opening statements will begin in the defamation lawsuit against television talk show host Oprah Winfrey. Some cattle ranchers say comments she made on her show about mad cow disease caused beef prices to fall. Today, a jury was selected.

Fig. 3 Document content of the three approaches

Table 2. Types of Terms in Setup 3

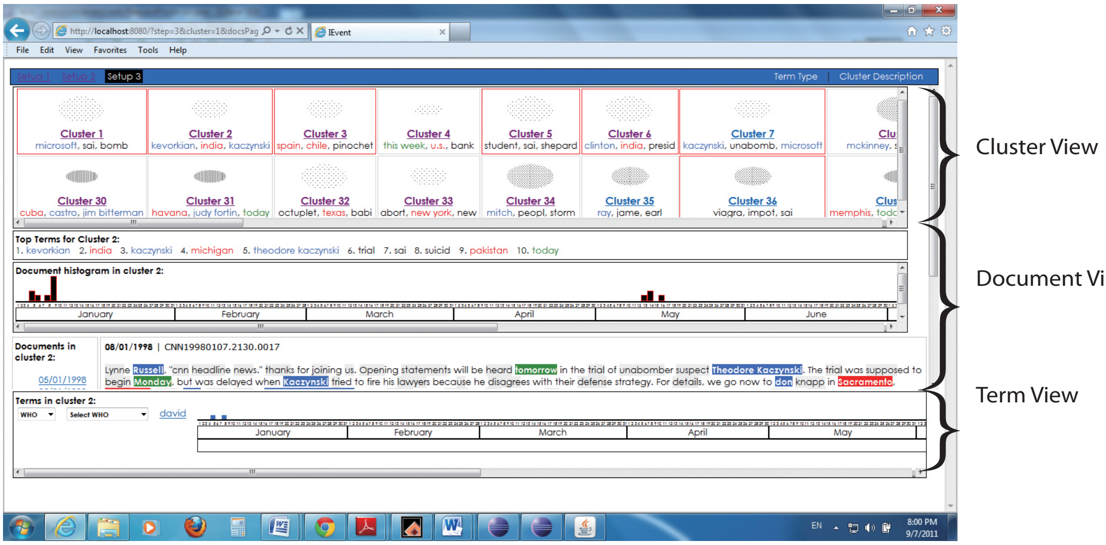
Type of Term	Example
WHO: person, organisation	Oprah, CNN
WHERE: location	Texas, U.S.
WHEN: date	April, Friday, tomorrow, 1996
WHAT: keyword	Oprah Winfrey, cow, beef, disease

3.3. Term View (TV)

Term View provided information on the occurrence of keywords and NE within the timeline with a histogram in a cluster, and an option terms list where there is no difference in timeline across setups. The difference only appeared in the option terms list, where Setup 1 (baseline setup) provided the user with "What" only and Setup 2 (baseline setup) provided the user with "Who, Where, and When"; meanwhile Setup 3 (experimental setup) provided the user with "Who, Where, When, and What."

The three components were arranged on the interface from top to bottom based on their relevance to the user, beginning with Cluster View (CV), followed by Document View (DV), and lastly Term View (TV). CV enabled the users to browse the whole collection before they went on to search a specific cluster, while DV allowed them to view all the documents in the chosen cluster and provided them with an interactive graphical timeline interface to browse the major events and the cluster contents. Finally, TV displayed the terms in the cluster and provided a timeline to link the terms with the related document in DV. The components were arranged in an inverted pyramid on the interface to help users to narrow their browsing and to be more focused in their searching. In this study, *i*Event had three setups. Setup 1 and Setup 2 (Mohd et al., 2012) are the baseline setups and Setup 3 (Fig. 4) is the experimental setup.

*i*Event Setup 3 provided users with the same amount of data as in Setup 1 and Setup 2 with an equal number of documents and clusters on the interfaces. Setup 1 (baseline Setup) and Setup 2 (baseline Setup) used the same interface components and features as Setup 3 (experimental Setup), but differed in using either keywords or NE respectively instead of a combination approach (BOW+NE). Therefore, the user had the option to choose which Setup to use and the option to highlight the named entities on the interface using Setup 2 and Setup 3.


http://localhost:8080/?step=3&cluster=1&docsPag
IEvent
File Edit View Favorites Tools Help
Setup 1
Setup 2
Setup 3
Term Type | Cluster Description
Cluster 1 microsoft, sai, bomb
Cluster 2 kevorkian, india, kaczynski
Cluster 3 spain, chile, pinochet
Cluster 4 this week, u.s., bank
Cluster 5 student, sai, shepard
Cluster 6 clinton, india, presid
Cluster 7 kaczynski, unabomb, microsoft
Clu mckinney,
Cluster 30 cuba, castro, jim bitterman
Cluster 31 havana, judy fortin, today
Cluster 32 octuplet, texas, babi
Cluster 33 abort, new york, new
Cluster 34 mitch, peopl, storm
Cluster 35 ray, jame, earl
Cluster 36 viagra, impot, sai
Clus memphis, todc
Document histogram in cluster 2:
January
February
March
April
May
June
Documents in cluster 2:
05/01/1998
Terms in cluster 2:
WHO
Select WHO
david
Cluster View
Document View
Term View
EN
8:00 PM
9/7/2011

Top Terms for Cluster 2:
1. kevorkian 2. india 3. kaczynski 4. michigan 5. theodore kaczynski 6. trial 7. sai 8. suicid 9. pakistan 10. today

08/01/1998 | CNN19980107.2130.0017

Lynne Russell, "cnn headline news." thanks for joining us. Opening statements will be heard tomorrow in the trial of unabomber suspect Theodore Kaczynski. The trial was supposed to begin Monday, but was delayed when Kaczynski tried to fire his lawyers because he disagrees with their defense strategy. For details, we go now to don knapp in Sacramento.

Fig. 4 Bag of Words and Named Entities setup (BOW+NE)

4. USER EXPERIMENT

4.1. Experimental Data

We used a dataset that consisted of a collection of documents chosen from TDT2 and TDT3 corpuses which contained 1,468 CNN news documents. We used a copy of the dataset from the *i*Event experiment (Mohd et al., 2012) which involved 28 topics from the TDT2 corpus and 25 topics from the TDT3 corpus. Generally, the number of documents chosen will be adequate for the experiment where the collection's size is enough to produce clusters to be displayed on the user interface. CNN news stories were chosen because the stories from this source are mostly short, and modelling such stories is, we believe, a good first step before dealing with more complex datasets. Moreover, the dataset used was from 1998 as mentioned above, because users would have no familiarity with or interest in the topics during the evaluation of *i*Event.

4.2. Evaluation Users

The users were a mixture of journalists and postgraduate journalism students at the Universiti Kebangsaan Malaysia. Ten users were chosen, of whom three were female. 50% of the respondents were journalists and the remaining proportion were journalism students.

4.3. Experiment Evaluation

The evaluation of text classification effectiveness in an IR experimental system is often based on statistical measurements. These measurements are precision, recall, and F1-measure (Rijsbergen, 1979; Kowalski, 1997). In our case, the *retrieved* documents indicated those that had been classified by the system as positive instances of an event and the *relevant* documents were those that had been manually judged to be related to an event.

We treated each cluster as if it were the desired set of documents for a topic. The selection of topics for the Tracking in this experiment were based on the F1-measure (Mohd et al., 2012; Lewis & Gale, 1994; Pons-Porrata et al., 2004). To validate whether the *i*Event interface with combined BOW+NE helps the user to perform the TDT tasks, users were given a bad cluster to track or a bad topic to detect, based on the F1-measure. The chosen topics and clusters given in the user experiment had a combination of good and poor (94.7% - 50.9%) clustering performance according to F1-measure (Mohd et al., 2012).

4.4. User Tasks with *i*Event

There were two tasks, Tracking and Detection, which were given to the users.

4.4.1. Tracking Task

The definition of the Tracking task was to track the cluster that contained the identified topic. The user had to track the cluster that included the given topic and show that a sufficient amount of information on the event was provided by the system. This is consistent with the mission of a news journalist. Reporting and Profiling are sub-activities of this task where Reporting is defined as writing an article about a given topic; it requires the user to write an article about a topic through the formulation of significant facts. Meanwhile, Profiling is defined as drafting significant keywords as an outline for a topic. These tasks were considered to simulate natural news monitoring behavior by users. We aimed to create a type of interaction between the users and *i*Event where the latter could perform their own daily news tasks, such as everyday news monitoring and reporting tasks. To facilitate this, we placed the tasks within simulated cases (Borlund, 2002; Borlund & Ingwersen, 1997) shown in Fig. 5, where an example is given of a simulated case in a Topic Tracking task, since the given task's scenarios should reflect and support a real news monitoring situation.

SIMULATED SITUATION

It is early Spring 1998 in Washington, D.C. The last-minute negotiations Friday between tobacco companies (Philip Morris Inc., R.J. Reynolds Tobacco Co., Brown & Williamson Tobacco Co., Lorillard Tobacco Co., and United States Tobacco Co.) and Texas officials paid off in the form of a $15.3 billion settlement. Texas Attorney General Dan Morales said the deal would be of particular benefit to the state's children and taxpayers who have footed extra health care costs. The debate and activity of devising a National Tobacco Company Settlement seemed comfortable for everyone.

You have been asked to write an article on the outcome of the National Tobacco Settlement.

This topic is likely to generate events such as Clinton works with Congress on tobacco legislation, internal memos threaten national tobacco settlement, individual state tobacco settlements, hearings about national tobacco settlement, tobacco company sells high-nicotine cigarettes overseas, White House presses for national settlement, additional tobacco memos released, and budget would put tobacco settlement money into Medicare.

Fig. 5 Simulated situation

The steps in carrying out the Tracking task were:

1. A topic summary was given to the users.
2. Based on the information contained in the topic summary, the users were asked to track the related cluster.
3. The users investigated the documents in the related cluster they found.
4. The users were asked to draft the important facts or points to execute the Reporting task.
5. They listed out the dealing cluster.
6. The users created a profile by drafting useful keywords to perform the profiling task.
7. The users expressed their opinion on the features provided by *i*Event when performing the Tracking task by completing a questionnaire.

4.4.2. Detection Task

The detection task was the second task performed by the users. It is defined as identifying the topic dealt by a definite cluster. This is consistent with a journalist's task of identifying significant events that happened on a definite day (Mohd et al., 2012). The steps to carry out the Detection task are:

1. A specific cluster was given to the users with a list of twenty topics.
2. They were then asked to detect the topics from the documents contained in a specific cluster using any features of *i*Event to perform this task.
3. The users were asked to rank a maximum of three topics from the list of twenty topics given, if they felt that the specific cluster contained more than one.
4. Finally, they expressed their opinion on the features provided by *i*Event whilst performing the Detection task by completing a questionnaire.

In the Tracking task, there were six topics used and the users were given 15 minutes to complete each topic. They were given a total of 1:30 hours to attempt the entire Tracking task in three sessions.

In the Detection task there were also six clusters given where the users spent a maximum of 10 minutes to complete each cluster. In total, they were given one hour to complete the task in three sessions. Finally, at the conclusion of the tasks, the users completed a questionnaire which took approximately 5 minutes. The whole user experiment took about 2 hours 30 minutes to 3 hours excluding a short training session.

Users were given an opportunity to carry out the tasks using the interface. A Latin Square design (Spärck-Jones, 1981; Spärck-Jones & Willett, 1997; Doyle, 1975) was used which allowed us to evaluate the same topic using different setups. Fig. 6 shows the experimental design, where 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 factor.

Topic 1 (Oprah Lawsuit), for example, has an opportunity to change sequentially from first until sixth in order during the Tracking task. The clusters assigned in the Detection task were hidden during execution of the Tracking task to avoid any intersection of clusters that may have affected users' performance. The clusters were completed using single pass clustering with the threshold *t*=1.48 which created 57 clusters.

5. RESULTS

5.1. General Findings

During the experiment, 120 tasks were performed. 50% of these tasks were Tracking while the remaining tasks were Detection. The general results showed that 90% of the users liked to use *i*Event in both tasks and 10% of users disliked *i*Event because they were accustomed to other news tools. Those who disliked *i*Event 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 Bernama.com.

The results also revealed that more than 70% of the users that used *i*Event in the Tracking task found that the combination approach (BOW+NE) in Setup 3 provided more assistance in conducting the task compared with approximately 17% who found that the keywords in Setup 1 provided more assistance and 13% who found that named entities in Setup 2 provided more assistance. Meanwhile, 90% of all users considered that Setup 3 facilitated the Detection task. During the post-study interviews the users confirmed that the use of a combination of BOW and NE would help journalists to conduct their tasks since it provides significant information on the Who, Where, When, and What of an event at the same time.

A possible explanation for these results might be the users' success in performing both tasks (see Fig. 7). It is clear that there was a change in the users' topic interest after using *i*Event across setups. The Kruskal-Wallis Test confirmed that there was no statistically significant difference (*p*=0.742) in topic interest before using *i*Event across setups. Meanwhile, there was a statistically significant difference (*p*=0.010) in topic interest after using *i*Event across setups. The users were more interested in a topic in the Tracking task after using Setup 3 (*mean*=4.05 *sd*=0.686) as shown in Fig. 7.

Users	Session 1				Session 2				Session 3			
	Tracking		Detection		Tracking		Detection		Tracking		Detection	
	T1	T2	D1	D2	T3	T4	D3	D4	T5	T6	D5	D6
1- 3	**S1**	**S1**	S1	S1	S2	S2	S2	S2	S3	S3	S3	S3
4-6	S2	S2	S2	S2	S3	S3	S3	S3	**S1**	**S1**	S1	S1
7-10	S3	S3	S3	S3	**S1**	**S1**	S1	S1	S2	S2	S2	S2

*S1=Setup 1 (baseline setup); S2=Setup 2 (baseline setup); S3=Setup 3 (experimental setup)

Fig. 6 Experimental Design

During the *i*Event post-evaluation, the users 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 the combination approach helped to facilitate the user in all tasks as shown in Fig. 8. In the interview, the users told us that the new approach provided them with high quality forms of information where it was more descriptive. This contradicts findings by Mohd et al. (2012) which reported that the BOW approach provided better explanation in the Detection task and the NE approach in the Tracking task. However, in our findings users preferred both approaches to be offered in the Tracking and Detection task since they complement each other. Users receive specific information and at the same time are not missing any relevant information regarding a topic. Thus Setup 3 was significantly helpful in both tasks (Tracking and Detection) because it provided users with broad information (combinations from BOW and NE).

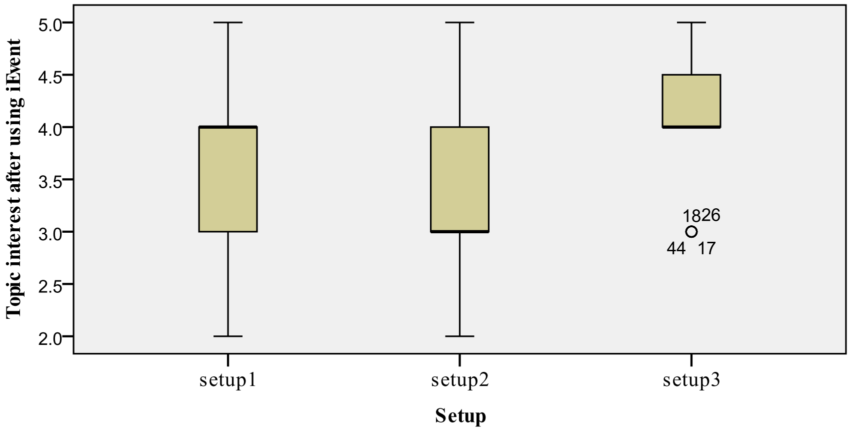


Fig. 7 Box plot of users' topic interest (scale 1-5) after using *i*Event across setups in the Tracking Task

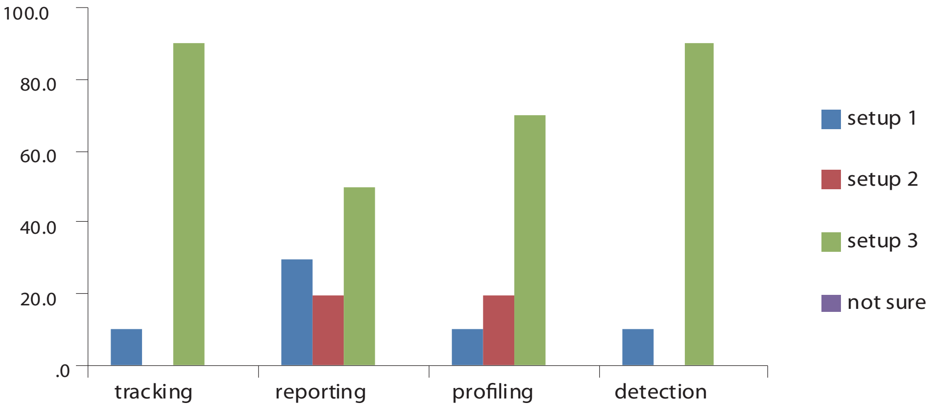


Fig. 8 Percentage of users' preferred setup for the given tasks

5.2. Tracking Task

First, the users' overall opinions on the use of the combination approach (BOW+NE) in Setup 3 were examined. Next, we investigated the users' 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, NE, or a combination). Finally, we investigated whether the users agreed that the use of the combination approach in *i*Event was perceived as useful, effective, helpful, and interesting in performing the TDT tasks.

5.2.1. Overall Opinions

Users' opinions of *i*Event during the Tracking task were analysed between setups. We investigated whether they perceived *i*Event as easy, relaxing, simple, satisfying, and interesting across setups. The Kruskal-Wallis Test confirmed that there was a statistically significant difference in users' opinion of easy (*p*=0.027) in conjunction with the setups. However, there was no statistically significant difference in simple (*p*=0.111), satisfying (*p*=0.390), relaxing (*p*=0.314), and interesting (*p*=0.155). Setup 3 was significantly stronger in easy (scale 4, 5) than Setup 1 and Setup 2 during the Tracking task, as appears in Table 3. Scale 1, 2, and 3 are treated as negative due to the zero value that appeared in scale 1 (Linacre, 2006). Meanwhile, scale 4 and 5 are considered as positive.

Easy

A ratio of 9 users to 1 found that Setup 3 was easier than the other Setups. 65% of users agreed that Setup 3 (*mean*=4.15 *sd*=0.587) was easy (scale 4) and 25% of them gave scale 5. Interestingly, none of the users found it difficult. This indicates that the use of the combination approach made the Tracking task easier.

Table 3. Summary Percentage of Users' Opinions (easy, relaxing, simple, satisfying, and interesting) Across Setup in Tracking Task

	Scale (%)					**(%)**		
	1	2	3	4	5	(-)ive	(+)ive	Ratio
Easy								
Setup1	0	5	30	55	10	35	65	2:1
Setup2	0	10	30	45	15	40	60	3:1
Setup3	0	0	10	65	25	10	90	**9:1**
Relaxing								
Setup1	0	10	25	60	15	35	75	2:1
Setup2	0	15	30	35	20	45	55	1:1
Setup3	0	5	10	65	20	15	85	**6:1**
Simple								
Setup1	0	15	30	35	20	45	55	1:0
Setup2	0	15	45	40	0	60	40	2:1
Setup3	0	5	30	40	25	35	65	2:1
Satisfying								
Setup1	0	15	5	60	20	20	80	4:1
Setup2	0	10	30	45	15	40	60	3:1
Setup3	0	0	20	60	20	20	80	4:1
Interesting								
Setup1	0	5	30	45	20	35	65	2:1
Setup2	0	5	30	55	10	35	65	2:1
Setup3	0	0	15	55	30	15	85	**6:1**

(-) ive =scale 1, 2,3; (+) ive = scale 4, 5
(Scale from 1 to 5, higher=better; highest value shown in bold)

Relaxing, Simple, Satisfying, and Interesting

As mentioned above, the Kruskal-Wallis Test confirmed that there was no statistically significant difference (p>0.05) in users' opinions with regard to relaxing, satisfying, and interesting. It clearly appears, however, that Setup 3 scored higher in terms of scale and ratio compared with Setup 1 and Setup 2, as shown in Table 3.

In the interview session, the users agreed that the use of the combination approach (BOW+NE) provided them with good findings and quick and precise information on the Who, Where, When, and What of an event, which made the Tracking task easy using Setup 3.

Interestingly, the satisfaction factor was also related to the high percentage of correct clusters tracked. This provides strong evidence that *i*Event helped significantly to facilitate users in tracking the correct cluster (*mean*=3.98 *sd*=0.129). The Kruskal-Wallis Test also confirmed that there was no statistically significant difference in the number of correct clusters to be tracked (*p*=0.368) in conjunction with the setups. This indicates that the users managed to track the correct clusters using all setups. We classified the correctness of cluster as tracked into four categories:

a. None - where users did not provide any information or they did not complete the task.
b. Wrong - where users tracked the wrong cluster.
c. Partially Correct - where users list out the minor cluster as their main finding.
d. Correct - where users list out the major cluster as their main finding.

The complete Tracking task was successful with 98.33% of tasks as correct and 1.67% as partially correct because sometimes a wrong cluster was chosen that caused wasted time in performance of the task (15 minutes). Moreover, users were confused by some terms, e.g. Merge, Texas, in topics which were highlighted in a cluster on a wrong topic.

The average time taken to conduct this task successfully across setups was approximately 12 minutes, 13 seconds. The users spent about 11 minutes, 35 seconds using the combination approach (BOW+NE) in Setup 3, 11 minutes, 20 seconds using NE in Setup 2, and 13 minutes, 45 seconds for BOW in Setup 1. The time taken in Setup 3 represents a mid-point between the shorter time that was taken in Setup 2 and the longer time in Setup 1. A possible explanation for the mid-time taken in Setup 3 was that it represented a compromise between the specific information (NE) provided by Setup 2 and the general information provided by Setup 1 (BOW). Since Setup 3 provided the users with a meaningful information, both tasks are facilitated.

5.2.2. Reporting Task

The Reporting task is one of the sub activities in Tracking. In this section, we report the findings of users' performance during this task after analysis of the number of lines that users wrote across setups. There was no statistically significant difference in the amount of news written in conjunction with the setups (Kruskal-Wallis Test, *p*=0.536), as shown in Fig. 9. The users managed to write almost the same number of lines in the report on average using Setup 1 (*mean*= 7.50 *sd*=2.819), Setup 2 (*mean*=6.45 *sd*=2.605) and Setup 3 (*mean*=7.45 *sd*=2.762).

5.2.3. Profiling Task

The users were asked to write important keywords as a profile for a topic where there were three types of keywords: named entities, terms, and combination (terms and named entities). For example, users might write keywords such as Exxon (NE), merge two companies (terms), and Exxon merge (combination) for the topic Mobil-Exxon Merger. We analysed the types of keywords provided by the users. The Kruskal-Wallis Test confirmed that there was no statistically significant difference between the types of keywords and the setups used in the Profiling task (*p*=0.504).

Findings revealed that users had a tendency to provide more terms when they were using Setup 1 and larger named entities when using Setup 2. When they were provided with a combination of BOW and NE in Setup 3, however, the results indicated that they prefer to use NE as the important terms in the profile of the task. Fig. 10 shows that Setup 3 provided users with a larger amount of information (34.74%) compared with Setup 1 (32.83%) and Setup 2 (31.93%). The users interviewed agreed that the new approach (BOW+NE) provided a high quality source of information (balance distribution of terms and named entities), thus it was suitable for use in the Profiling task.

We can conclude, therefore, that the combination of BOW and NE is a good approach in this task. Users are provided with broad terms, and at the same time the combination automatically provides them with highly significant terms (NE).

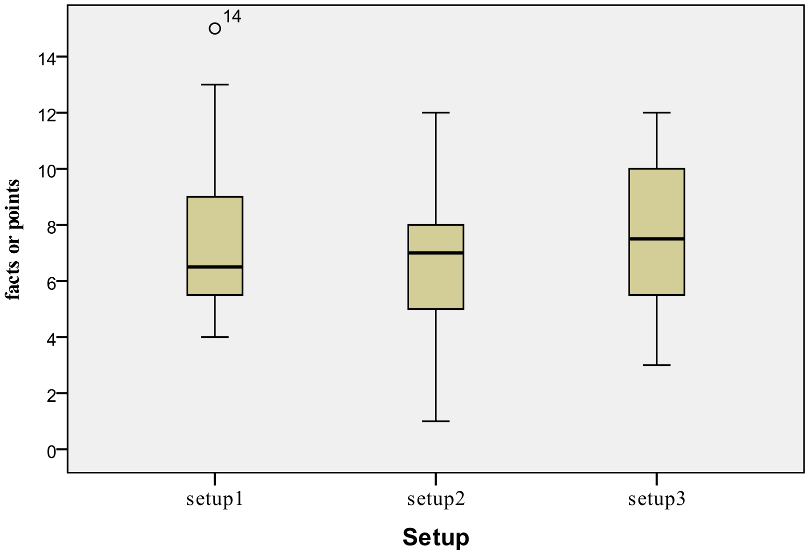


Fig. 9 Amount of lines written

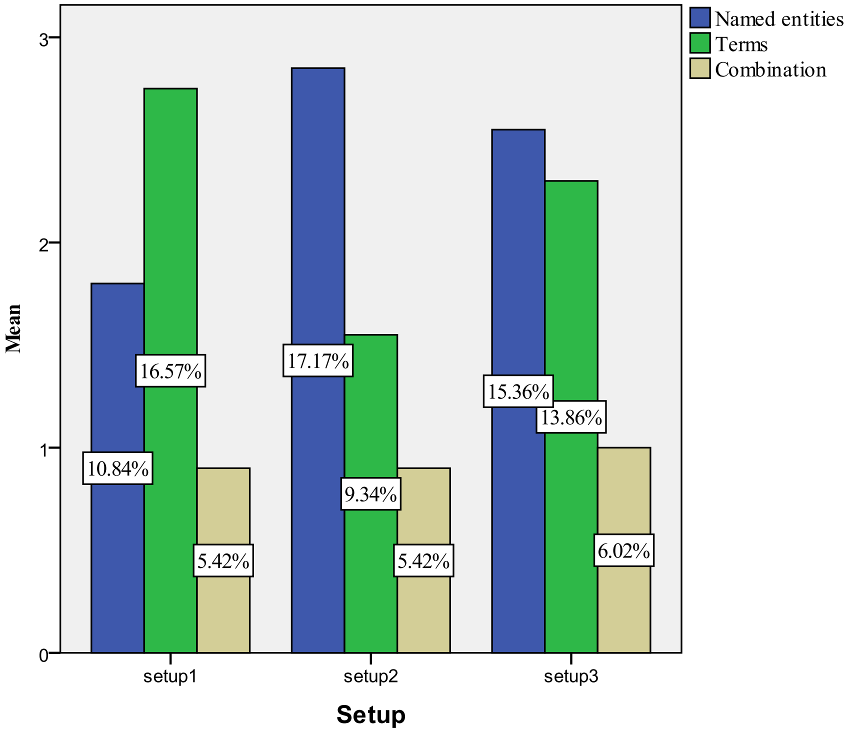


Fig. 10 Percentage for types of keywords across setups

5.2.4. Features

This section deals with analysing each feature of *i*Event across setups that users perceived as being useful, effective, helpful, and interesting during the Tracking task.

Useful

There was a statistically significant difference shown in the evaluation of the 'CV: cluster labelling' feature across setups (Kruskal-Wallis Test, *p*=0.030). The 'CV: cluster la-belling' feature in Setup 3 of *i*Event was more useful (*mean*=4.15 *sd*=0.875) compared to Setup 1 (*mean*=3.85 *sd*=1.137) and Setup 2 (*mean*=3.30 *sd*=1.031). It is clear from Table 4 that users found the 'CV: cluster labelling' feature useful when they were using Setup 3.

A ratio of 9 users to 1 found that the combination approach (BOW + NE) that was used in the 'CV: cluster labelling' feature in Setup 3 of *i*Event was significantly useful with 35% and 55% of users agreeing that it was useful (scale 4, 5), respectively as it seems to show in Table 4.

We can conclude that users found the combination approach (BOW+NE) used in the 'CV: cluster labelling' features in Setup 3 of *i*Event more useful than other approaches used in Setup 1 and Setup 2.

Effective

There was also a statistically significant difference shown in the evaluation of the 'CV: keyword approach' feature across setups (Kruskal-Wallis Test, *p*=0.049). Setup 2 was more effective (*mean*=4.20 *sd*=0.523) compared to Setup 3 (*mean*=4.50 *sd*=0.668) and Setup 1 (*mean*=3.80 *sd*=1.105). Users found this feature in Setup 2 and 3 of *i*Event more effective as shown in Fig. 11.

Table 4. 'CV: cluster labelling' Feature Across Setups Perceived as Useful

FEATURES	Scale (%)					(%)		
	1	2	3	4	5	(-)ive	(+)ive	Ratio
CV: cluster labelling								
Setup 1	0	20	10	35	35	20	70	4:1
Setup 2	0	30	20	40	10	30	50	2:1
Setup 3	0	10	0	55	35	10	90	**9:1**

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

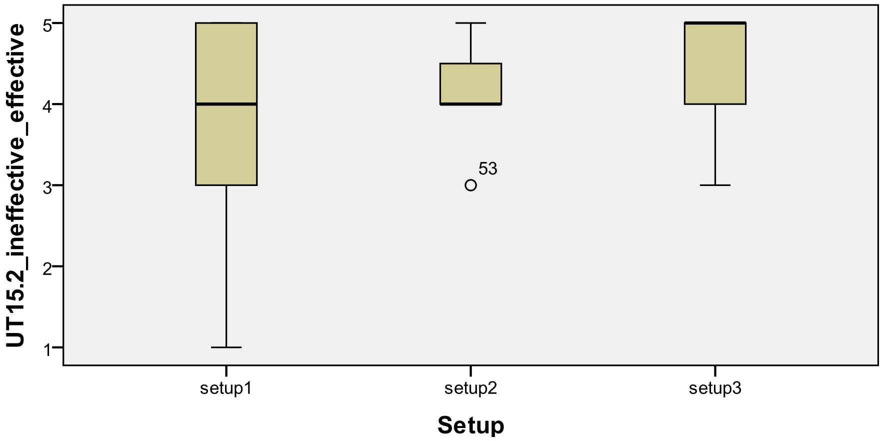


Fig. 11 Box plot on effectiveness (scale 1-5) of the 'CV: keyword approach' feature across Setups

Helpful

For this opinion value, the Kruskal-Wallis Test also confirmed that there was a statistically significant difference across setups shown in the evaluation of the 'CV: cluster labelling' feature (p=0.011, with *mean*= 4.30 *sd*=0.923), in that Setup 3 of *i*Event was more helpful than Setup 1 (*mean*=3.75 *sd*=1.251) and Setup 2 (*mean*=3.25 *sd*=1.020). Users found this feature was helpful when they were using Setup 3, as shown in Table 5. A ratio of 16 users to 1 found that the combination approach (BOW+NE) used in the 'CV: cluster labelling' feature in Setup 3 of *i*Event was significantly helpful, with 55% of users agreeing that it was very helpful (scale 5).

This indicates that users found the combination approach that was used in the 'CV: cluster labelling' features in Setup 3 of *i*Event more helpful than Setup 1 and Setup 2.

Interesting

For this opinion, the Kruscal-Wallis Test confirmed there was a statistically significant difference (p<0.05) on two features across setups. These were the 'CV: top terms' feature (p=0.002) and 'CV: cluster labelling' feature (p=0.022).

The evaluation shows that all features of *i*Event were perceived as being interesting in Setup 3 (scale 4, 5) compared to Setup 1 and Setup 2, as shown in Fig. 12. These are 'CV: top terms' (*mean*=4 *sd*=0973) and 'CV: cluster labelling' (*mean*=4.25 *sd*=0.851) features.

Table 5. 'CV: cluster labelling' Features Across Setups Perceived as Helpful

FEATURES	Scale (%)					(%)		
	1	2	3	4	5	(-)ive	(+)ive	Ratio
Setup 1	5	15	15	30	35	20	65	3:1
Setup 2	0	25	40	20	15	25	35	1:1
Setup 3	0	5	15	25	55	5	80	16:1

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

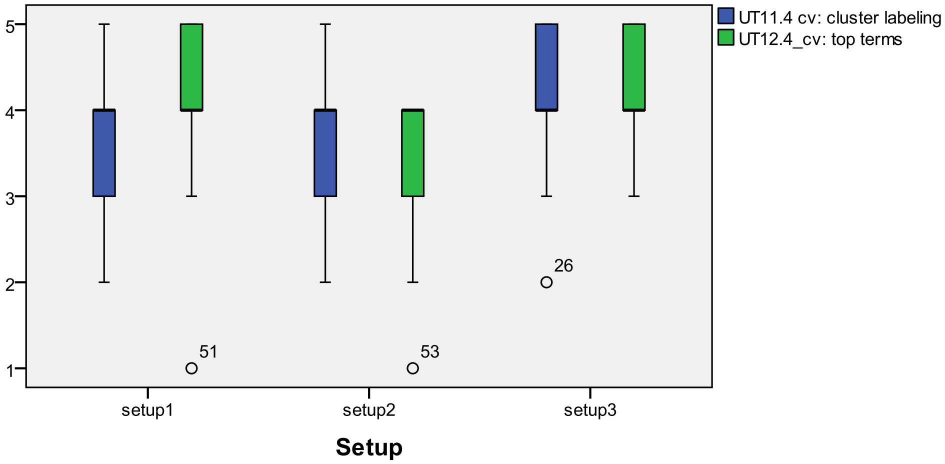


Fig. 12 Box plot on Interestingness (scale 1-5) of 'CV: cluster labelling', 'CV: top terms' features across Setups

5.3. Detection Task

This section is concerned with the analysis of users' agreement on the ease of detecting the topics and frequency of features with the combination approach (BOW+NE) across setups.

The whole Detection task was conducted successfully with 83.33% of task findings being correct and 13.33% being partially correct. There were 3.33% unsuccessful Detection tasks; we believe the reason for this is because of some users' confusion between the topics. Generally, this confirmed that *i*Event managed to facilitate user performance in the Detection task, as shown in Fig. 13.

The Kruskal-Wallis Test also confirmed that there was no statistically significant difference in the number of correct topics detected (*p*=0.668) across the setups. This proves that the users managed to detect the correct topics using all setups. We classify the correctness of topics detected into four categories:

a. None - where users did not gather any information or they did not complete the task.
b. Wrong - where users detected the wrong topic.
c. Partially Correct - where users listed out the minor topic as their main finding
d. Correct - where users listed out the major topic as their main finding

It was apparent from the Kruskal-Wallis Test that there was no statistically significant difference between the users' opinions (*p*=0.369) on the ease of detecting a topic using setups. However, it seems from Fig. 14 that it was easier to detect a topic using Setup 3 than with Setup 1 and Setup 2.

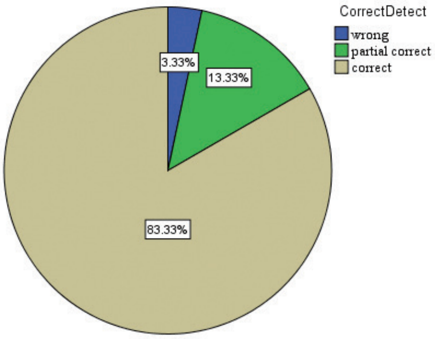


Fig. 13 Percentage of successful Detection tasks

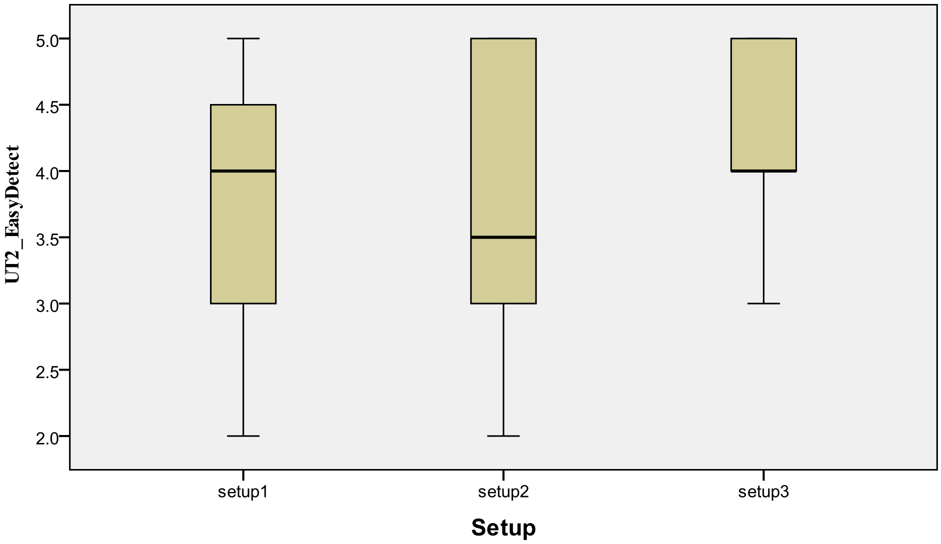


Fig. 14 Box plot on Easy to Detect (scale 1-5) with the Setup for the Detection task

As mentioned previously in Section 5.1 (General Findings), during the post-evaluation interview the users found that the combination approach (BOW+NE) was more descriptive because they had to detect the topic dealt by a specific cluster in the Detection task, and Setup 3 (combination approach) provided users with Who, Where, When, and What information at the same time.

As shown in Table 6, 55% of users agreed that it was easy (scale 4) to detect a topic using Setup 3 (*mean*=4.40, *sd*=0.699). Surprisingly, none of the users found that it was hard to detect a topic using any of the setups. 85% of users found that Setup 3 makes the Detection task easier than Setup 1 and Setup 2.

Further analysis on the interaction logs proved that users had a higher activity for the 'DV: document content' feature in Setup 3 (77.3%) compared to Setup 1 (68.8%) and Setup 2 (65.7%). These results clarified that users clicked less on the 'DV: document content' feature with NE (Setup 2) and BOW (Setup 1), than with the combination approach (BOW+NE) in Setup 3. Highlighting BOW and NE in the 'DV: document content' feature has provided users with quick and significant information rather than displaying plain document content. In addition, it provided users with links to highlight other clusters which had the NE or BOW in the document content by clicking on the terms or named entities. Additionally, in the successful attempts at the Detection task, the users took less time using the combination approach (BOW+NE) (00:04:22) compared to using NE (00:04:41) and keywords (00:04:50).

Furthermore, the interaction logs showed that there was slight activity in the 'TV: keyword approach' feature. This was seen clearly in Setup 3 and less clearly in the other two Setups. This may be due to users attempting to use Who, Where, When, and What terms.

These results indicate that the combination approach (BOW+NE) used in the 'DV: document content' feature provided them with all the terms, and the NE used in the 'DV: document content' feature helps them to make a quick decision in respect to a topic in the Detection task.

6. DISCUSSION

In general, most of the features in Setup 2 (NE) facilitated users in the Tracking task. Meanwhile, Setup 1 (BOW) facilitated them in the Detection task. Interestingly, users agreed that most of the features in the combination approach (BOW+NE) facilitated performance of both tasks. Users found, however, that the use of the combination approach (BOW+NE), for example in the 'DV: document content' feature, also provided them with broader information that helped them to detect the topic quickly. This proves that the combination approach (BOW+NE) was effective and efficient in both tasks. Table 7 shows the features with their approaches which facilitated users in performing the TDT tasks.

The Kruskal-Wallis Test confirmed that there was no statistically significant difference between the successful Tracking (p=0.368) and Detection task (p=0.668) across setups. It is seems clear that there is consensus that the combination approach (BOW+NE) helped the users in both tasks; moreover, most of them agreed that it was useful and interesting.

Future analysis on the interaction logs enabled us to make a comparison of the mean number of clicks among the successful tasks between setups, as shown in Table 8.

It seems from Table 8 that there is a disparity in mean number of clicks between the approaches that were used in performing tasks across setups. This can be seen clearly between Setup 1 and Setup 2 where mean clicks in Setup 2 in the Tracking task are higher than for Setup

Table 6. Percentage of User Opinions on Ease of Detecting a Topic Across Setups

	Scale (%)					**(%)**	
	1	2	3	4	5	(-)ive	(+)ive
Setup 1	0.0	10	25	40	25	10	65
Setup 2	0.0	15	35	15	35	15	50
Setup 3	0.0	0	15	55	30	0	85

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

Table 7. Comparison of Features, with Approaches, Which Facilitated TDT tasks

	Keywords	Named entities	BOWs+NE
CLUSTER VIEW (CV)			
cluster labelling	Detection	Tracking	Both
top terms	Detection	Tracking	Both
DOCUMENT VIEW (DV)			
document content		Tracking	Both
TERM VIEW (TV)			
keyword approach	Detection	Tracking	Both

Both= Tracking+Detection tasks

Table 8. Comparison of Features, with Approaches, Which Facilitated TDT tasks

Task	Mean of click		
	Setup 1	Setup 2	Setup 3
Tracking	23	**37**	31
Detection	**32**	25	28

(Highest value shown in bold)

1. Meanwhile, the mean number of clicks in Setup 1 is higher in the Detection task. Surprisingly, it is clear that the mean clicks in Setup 3 was closely equal to the mean average of each task in both Setup 1 and Setup 2, showing that the combination approach (BOW+NE) facilitated both tasks. There are two explanations for the previous findings in interaction logs: First, the higher number of clicks indicates a lack of information or higher interaction between the users and the system; second, the average number in clicks provided by Setup 3 may indicate that users are provided with enough information to perform both tasks.

7. CONCLUSIONS

Users are provided with the two types of information and they can choose to display either keywords or keywords with named entities 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. Thus, this study has presented an experimental study conducted with journalists to investigate the combination of bag of words and named entities (BOW+NE) approaches implemented in the *i*TDT interface called the Interactive Event Tracking (*i*Event) system; this includes what TDT tasks these approaches facilitate. The combination approach (BOW+NE) in Setup 3 has provided the users with focused information to facilitate both tasks and enhanced interaction between users and the system. Users provided a large number of terms (BOW and NE) during the Profiling task using Setup 3, which indicates that the *i*TDT with the combination approach (BOW+NE) is useful in helping journalists to perform their tasks. It has provided users with descriptive and significant information. Overall these findings revealed that the use of a combination approach (BOW+NE) effectively enhanced the *i*TDT interface and created more interaction between users and the system in both tasks (Tracking and Detection). Journalists require both descriptive information in the Detection task and significant information in the Tracking task. Therefore the combination approach (BOW+NE) in Setup 3 has provided users with focused information to facilitate both tasks and enhanced interaction between users and the system. As a further contribution, this article confirms the value of the combination approach in *i*TDT which has facilitated the journalists in performing the Tracking and Detection tasks.

REFERENCES

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., & Amstutz, P. (2005). Taking topic detection from evaluation to practice. *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05)* (pp. 101.1). Washington: IEEE Computer Society.

Berendt, B., & Subasic, I. (2009). STORIES in time: A graph-based interface for news tracking and discovery. *WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology,* (pp. 531- 534). Washington: IEEE Computer Society.

Borlund, P. (2002). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation, 53*(3), 225–250.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (pp. 168–175). Philadelphia: PA.

Doyle, L. B. (1975). *Information retrieval and processing.* California: Melville.

Jones, G. J. F., & Gabb, S. M. (2002). A visualisation tool for topic tracking analysis and development. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 389-390). New York: ACM Press.

Kowalski, G. (1997). *Information retrieval systems – Theory and implementation.* London: Kluwer Academic.

Kumaran, G., & Allan, J. (2004). Text classification and named entities for new event detection. *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 297–304). New York: ACM Press.

Leuski, A., & Allan, J. (2000). Lighthouse: Showing the way to relevant information. In *Proceedings of the IEEE Symposium on information Visualization* (pp. 125-129). Washington, IEEE Computer Society.

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-12). New York: Springer-Verlag.

Linacre, J. M. (2006). *WINSTEPS Rasch measurement computer program.* Chicago: Winsteps.

Makkonen, J., Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. *Information Retrieval, 7*(3-4), 347-368.

Mohd, M., Crestani, F., & Ruthven, I. (2012). Evaluation of an interactive topic detection and tracking interface. *Journal of Information Science, 38*(4), 383-398.

Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J., & Perez-Martinez, J. M. (2004). JERARTOP: A new topic detection system. In *Proceeding of Progress in Pattern Recognition, Image Analysis and Applications* (pp. 71-90). New York: Springer-Verlag.

Rijsbergen, C. J. (1979). *Information retrieval,* 2nd ed. London: Butterworths.

Spärck-Jones, K. S., & Willett, P. (1997). *Readings in information retrieval.* San-Francisco: Morgan Kaufmann.

Spärck-Jones, K. S. (1981). *Information retrieval experiment.* London: Butterworth-Heinemann Newton.

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 49-56). New York: ACM Press.

Yang, Y., Carbonell, J. Q., Brown R. D., Pierce, T., Archibald, B. T., & Lin, X. (1999). 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.

Zhang, K., Zi, J., & Wu, L. G. (2007). New event detection based on indexing-tree and named entities. *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 215- 222). Netherlands: ACM Press.