

5-2005

## Event-driven document selection for terrorism

Zhen SUN

Ee Peng LIM

Singapore Management University, eplim@smu.edu.sg

Kuiyu CHANG


Nanyang Technological University

Teng-Kwee ONG

Rohan Kumar Gunaratna

**DOI:** [https://doi.org/10.1007/11427995\\_4](https://doi.org/10.1007/11427995_4)

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

---

### Citation

SUN, Zhen; LIM, Ee Peng; CHANG, Kuiyu; ONG, Teng-Kwee; and Gunaratna, Rohan Kumar. Event-driven document selection for terrorism. (2005). *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2005, Atlanta, GA, USA, May 19-20, 2005: Proceedings*. 3495, 37-48. Research Collection School Of Information Systems.

**Available at:** [https://ink.library.smu.edu.sg/sis\\_research/890](https://ink.library.smu.edu.sg/sis_research/890)

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [libIR@smu.edu.sg](mailto:libIR@smu.edu.sg).

# Event-Driven Document Selection for Terrorism Information Extraction

Zhen Sun<sup>1</sup>, Ee-Peng Lim<sup>1</sup>, Kuiyu Chang<sup>1</sup>, Teng-Kwee Ong<sup>2</sup>,  
and Rohan Kumar Gunaratna<sup>2</sup>

<sup>1</sup> Centre for Advanced Information Systems, School of Computer Engineering,  
Nanyang Technological University, Singapore 639798, Singapore  
aseplim@ntu.edu.sg

<sup>2</sup> International Center for Political Violence and Terrorism Research,  
Institute of Defence and Strategic Studies,  
Nanyang Technological University, Singapore 639798, Singapore

**Abstract.** In this paper, we examine the task of extracting information about terrorism related events hidden in a large document collection. The task assumes that a terrorism related event can be described by a set of entity and relation instances. To reduce the amount of time and efforts in extracting these event related instances, one should ideally perform the task on the relevant documents only. We have therefore proposed some document selection strategies based on information extraction (IE) patterns. Each strategy attempts to select one document at a time such that the gain of event related instance information is maximized. Our IE-based document selection strategies assume that some IE patterns are given to extract event instances. We conducted some experiments for one terrorism related event. Experiments have shown that our proposed IE based document selection strategies work well in the extraction task for news collections of various size.

**Keywords:** Information extraction, document selection.

## 1 Introduction

### 1.1 Objectives

Information about a certain terrorism event frequently exists across several documents. These documents could originate from different portal and news websites around the world, with varying amount of information content. Clearly, it is not always necessary for a terrorism expert to find all documents related to the event since they may carry duplicate information. As far as the expert is concerned, it is important to read only a small subset that can give a complete and up-to-date picture so as to maximize his/her efficiency.

We therefore propose an extraction task, for which several pattern-based document selection strategies were studied. The extraction task aims to incrementally select a set of documents relevant to a terrorism event, using various

document selection strategies designed to aid in the extraction task. A set of patterns for extracting event specific entity and relationship instances from a document is assumed to be given. We also assume that some seed entity instances are given to bootstrap the extraction process. The overall objective is to find all entity and relationship instances related to the given event from the smallest possible subset of documents.

In the following, we summarize our contribution to this research:

- We formally define the extraction task which incorporate a document selection strategy to find event related entity and relationship instances. This task has not been studied before and our work therefore establishes the foundation for this field.
- We propose a few document selection strategies to identify the smallest possible subset of documents for event related instances. Each document selection strategy aims to maximize the novelty of the set of entity and relationship instances that can be found in the next document to be extracted. In this way, one can hopefully reduce the number of documents that the terrorism experts have to review.
- We have created two datasets to evaluate our extraction task and document selection strategies. The experimental results show that our strategies performs well and appears to scale well to large document collection.

## 1.2 Paper Outline

In the remaining portion of this paper, we present related work in Section 2, followed by formal definitions of the extraction task in Section 3. Next, we describe our proposed document selection strategies in Section 4. Experiments and results are given in Sections 5 and 6 respectively. Section 7 gives the conclusion and future work.

## 2 Related Work

Finding entity and relation instances of a certain event is our research focus and this is related to named entity recognition. Named Entity [4] Recognition deals with extracting specific classes of information (called "entities") such as person names, locations, and organizations from plain text. Michael Chau et al. [14] addressed the problem on extracting entities from police narrative reports. In their work, they built a neural network-based extraction method.

Named Entity Recognition can be viewed as a kind of single-slot extraction. Single-slot extraction such as AutoSlog [6] and its extensions [7, 8, 9] have been developed and which have demonstrated high extraction accuracy. Multi-slot extraction [2, 5, 10] refers to extracting multiple slots with some relationships from a document. Our work utilizes both multi-slot extraction and named entity recognition in the extraction task.

New Event Detection (NED) is a document selection task to identify the first story of an event of interest from an ordered collection of news articles. Since the

first story does not necessarily contain all the entity and relation instances of an event, NED methods cannot be applied directly to the event-driven extraction task [11, 12]. These methods often do not involve information extraction except in [13], where Wei and Lee proposed an information extraction based event detection (NEED) technique that uses both information extraction and text categorization methods to conduct NED.

Finn and Kushmerick proposed various active learning selection strategies to incrementally select documents from a large collection for user labelling so as to derive good extraction patterns [1]. In contrast, our work focuses on finding documents containing both novel and related information with the help of extraction patterns.

### 3 Event-Driven Extraction with Document Selection

#### 3.1 Event Representation Using Entity and Relation Instances

In our extraction task, we represent a terrorism event by a set of entity and relation instances. The entity instances describe the people, organisations, locations, dates/times and other information relevant to the event. The relation instances provide the links between entity instances so as to understand their inter-relations. Prior to the extraction task, we assume that a terrorism expert wishes to derive all entity and relation instances for a single event. To ensure that only relevant instances are extracted, a set of entity and relation classes are assumed to be known apriori.

Let  $E$  be a set of *entity classes*, i.e.  $E = \{E_1, E_2, \dots, E_n\}$ , and  $R$  be a set of *relation classes*,  $R = \{R_1, R_2, \dots, R_m\}$ .  $E$  and  $R$  together describe the information to be extracted for a target terrorism event. An entity class  $E_i$  denotes a set of entity instances of the same type, and each entity instance is usually a noun or noun phrase appearing in the document. Each relation class  $R_i$  represents a semantic relationship from an entity class  $SourceEnt(R_i)$  to an entity class  $TargetEnt(R_i)$  and is associated with an *action class*  $A_i$ .  $A_i$  refers to a set of verbs or verb phrases that relate source entity instances in  $SourceEnt(R_i)$  to target entity instances in  $TargetEnt(R_i)$ . Each relation instance comprises a source entity instance from  $SourceEnt(R_i)$ , a target entity instance from  $TargetEnt(R_i)$ , and an action instance from  $A_i$ , i.e.,  $R_i \subseteq SourceEnt(R_i) \times A_i \times TargetEnt(R_i)$ , where  $SourceEnt(R_i), TargetEnt(R_i) \in E$ .

#### 3.2 Event-Driven Extraction Task

Suppose we are given a set of extraction patterns  $EP$ , a collection of documents  $D$ , and a set of seed entity instances  $W$  relevant to an event. Let  $E$  and  $R$  represent the entity classes and relation classes relevant to the event. We use  $\mathcal{E}$  to denote the set of all entity instances contained in  $E$ , i.e.  $\mathcal{E} = \cup_{i=1}^n E_i$ , and  $\mathcal{R}$  to denote the set of all relation instances in  $R$  i.e.,  $\mathcal{R} = \cup_{i=1}^m R_i$ .  $W$  is a small subset of  $\mathcal{E}$  useful for bootstrapping the extraction of other instances. To ensure that all instances will be extracted given the seed entity instances  $W$ , we require

---

**Algorithm 1.** Event-Driven Extraction Task

---

**inputs:**  $EP, D, W$   
**for** each document  $d_j$  in  $D$  **do**  
  apply  $EP$  on  $d_j$  to obtain  $\mathcal{E}'_j, \mathcal{R}'_j$   
  score  $d_j$  using **InitialScore** function  $score(d_j)$   
**end for**  
**repeat**  
  Find the document  $d_s$  with highest  $score(d_s)$ , move  $d_s$  from  $D$  to  $S$   
  Extract (manually by an expert user) entity and relation instances from  $d_s$   
  Add newly extracted instances from  $d_s$  to  $\mathcal{E}$  and  $\mathcal{R}$   
  **for** each document  $d_j$  in  $D$  **do**  
    re-score  $d_j$  using a score function  $score(d_j)$  based on  $\mathcal{E}'_j, \mathcal{R}'_j, \mathcal{E}, \mathcal{R}$   
  **end for**  
**until** termination condition is satisfied  
**outputs:**  $S, \mathcal{E}, \mathcal{R}$

---

all event instances  $\mathcal{E}$  to be directly or indirectly linked to  $W$  through the relation instances in  $\mathcal{R}$ .

In the **event-driven extraction task**, documents for extracting event related instances are selected one at a time. At the beginning, the seed entity instances set  $W$  is given to identify the relevant documents. Each time a document is selected, it is given to the expert user for manual extraction of entity and relation instances. Note that manual extraction is conducted to ensure that no instances are missed. This process repeats until all event related entity and relation instances are found.

The detailed description of the task is depicted in Algorithm 1. During the extraction task, the extraction patterns  $EP$  are used to find the existence of entity and relation instances that could be relevant to the event. The extraction patterns can be for single-slot, or multi-slot extraction. The former is appropriate for extracting entity instances while the latter can be used for extracting both entity and relation instances. The entity and relation instances extracted from a document  $d_j$  using  $EP$  are stored in  $\mathcal{E}'_j$ 's and  $\mathcal{R}'_j$ 's respectively.

Assuming that the expert user has in mind a set of entity and relation instances to be extracted for an event. We can then define a set of documents containing relevant instances as the *relevant set* denoted by  $L$ . The objective of the event-driven extraction task on the other hand is to select the smallest subset  $O$  of  $L$  that covers all relevant instances. We call  $O$  the **optimal set**. Let  $\mathcal{E}_j$  and  $\mathcal{R}_j$  denote the set of entity and relation instances in document  $d_j$ . Then  $O$  is an optimal set if and only if it satisfies the following two conditions:

1.  $(\cup_{d_j \in O} \mathcal{R}_j = \mathcal{R})$  and  $(\cup_{d_j \in O} \mathcal{E}_j = \mathcal{E})$
2.  $\nexists O'$  s.t.  $(\cup_{d_j \in O'} \mathcal{R}_j = \mathcal{R})$  and  $(\cup_{d_j \in O'} \mathcal{E}_j = \mathcal{E})$  and  $(|O'| < |O|)$

## 4 Pattern-Based Document Selection Strategies

We have developed several document selection strategies using different score functions in the proposed extraction task. Each document selection strategy adopts a different score function to rank documents. In general, documents containing significant novel and related information should have higher scores. Since these strategies rely on extraction patterns to identify potentially relevant entity and relation instances, we call them *pattern-based document selection strategies*.

### 4.1 InitialScore

This is the default strategy that selects documents based on the given seed entity instances  $W$ . This document selection strategy is therefore used in the first iteration only. The primary objective of scoring is to assign higher scores to documents that have more extraction patterns fired. The first term of the score formula below considers proportion of  $W$  that is extracted. This is to ensure a relevant document is selected initially.

$$score(d_j) = \frac{|\mathcal{E}'_j \cap W| + \frac{|W|}{\gamma}}{|W|} \cdot \log_2(|EP_j|) \cdot \sum_{k=1}^{|EP_j|} f_k \quad (1)$$

where  $\gamma \gg |W|$  is a smoothing factor that prevents the first term from becoming zero if  $\mathcal{E}'_j \cap W$  is empty,  $EP_i$  is a subset of  $EP$  that fired on document  $d_j$ , and  $f_j$  is the number of relation instances extracted by extraction pattern  $ep_{j,k}$ . In our experiment, we used  $\gamma = 100$  with  $|W| = 4$ .

### 4.2 DiffCompare

This strategy examines the amount of overlap between relation instances extracted from the current document  $d_j$  with the accumulated relation instance set  $\mathcal{R}$ . The smaller the overlap, the higher the score. In addition, the amount of intersection between the extracted entity instances  $\mathcal{E}'_j$  and  $W$  is also considered. This is to assign higher score for documents having direct links to the seed set. Contribution from the two factors are linearly weighted by  $\alpha \in [0, 1]$ . Equation (2) shows the score function:

$$score(d_j) = \alpha \cdot \frac{|\mathcal{R}'_j - \mathcal{R}|}{\max_{d_i \in D} |\mathcal{R}'_i|} + (1 - \alpha) \cdot \frac{|\mathcal{E}'_j \cap W|}{|W|} \quad (2)$$

where  $N$  is the total number of documents in  $D$ .

### 4.3 CombineCompare

This strategy combines the amount of intersection and dissimilarity between relation instances extracted from  $d_i$  with instances in  $\mathcal{R}$ . A modifier  $\beta \in [0, 1]$  is used to adjust the relative importance of overlapping relation instances compared with novel relation instances (i.e., relevant relation instances that have not been

extracted so far). When the former is more important,  $\beta > 0.5$ . When  $\beta = 0.5$ , both are treated equally important. Equation (3) gives the score function of this strategy. Note that when  $\beta = 0$ , this is equivalent to DiffCompare.

$$\text{score}(d_j) = \alpha \cdot \frac{\beta \cdot |\mathcal{R}'_j \cap \mathcal{R}| + (1 - \beta) \cdot (|\mathcal{R}'_j - \mathcal{R}|)}{\max_{d_i \in D} |\mathcal{R}'_i|} + (1 - \alpha) \cdot \frac{|\mathcal{E}'_j \cap W|}{|W|} \quad (3)$$

#### 4.4 PartialMatch

In this document selection strategy, we want to select documents with relation instances linked to those entity instances that have already been found. This requires a partial match between the former and latter. Note that all entity instances in the event are connected with others using relation instances. This applies even in the midst of extraction task. Hence, we only need to conduct partial match between a relation instance extracted using *EP* and the relation instances found so far.

Given two relation instances  $r_s = (e_s^s, a_s, e_s^t)$  and  $r_t = (e_t^s, a_t, e_t^t)$ , the partial match of  $r_s$  and  $r_t$  denoted by  $PartialMatch(r_s, r_t)$  is defined by:

$$PartialMatch = \begin{cases} 0 & \text{if } e_s^s \neq e_t^s \wedge e_s^t \neq e_t^t \wedge (a_s \in A_p, a_t \in A_q, p \neq q) \\ 0 & \text{if } e_s^s = e_t^s \wedge e_s^t = e_t^t \wedge (a_s \in A_p, a_t \in A_q, p = q) \\ 0 & \text{if } e_s^s \neq e_t^s \wedge e_s^t \neq e_t^t \wedge (a_s \in A_p, a_t \in A_q, p = q) \\ 1 & \text{otherwise} \end{cases}$$

With  $PartialMatch$  measuring the novelty of instances, we now define the score function for the partial match document selection strategy in equation (4):

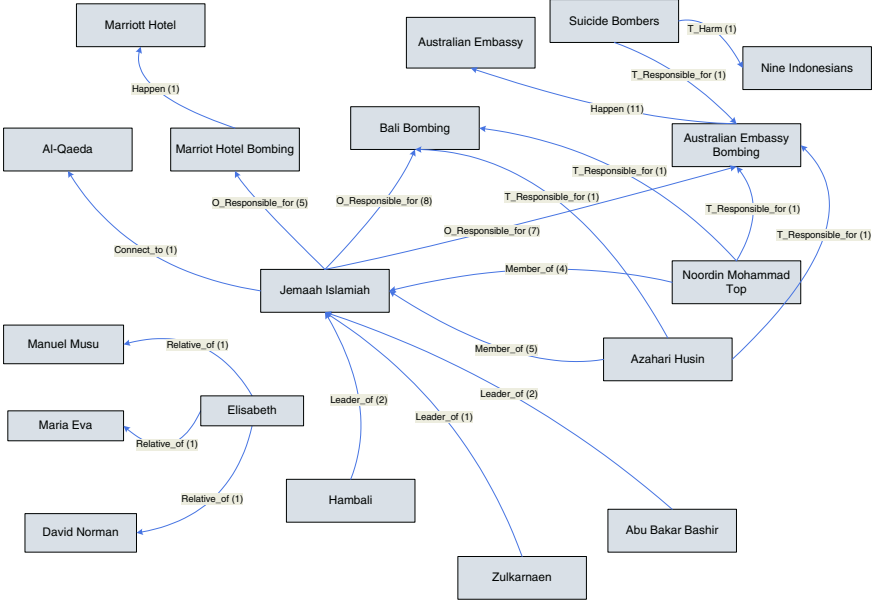
$$\text{score}(d_j) = \alpha \cdot \frac{\sum_{k=1}^{M_j} \sum_{h=1}^{|\mathcal{R}|} PartialMatch(r'_{j,k}, r_h)}{|\mathcal{R}'_j| \cdot |\mathcal{R}| + 1} + (1 - \alpha) \cdot \frac{|\mathcal{E}'_j \cap W|}{|W|} \quad (4)$$

where  $M_j$  is the number of relation instances extracted from  $d_j$  using *EP*;  $r'_{j,k}$  is the  $k$ th relation instance from  $\mathcal{R}'_j$ ; and  $r_h$  is the  $h$ th instance in  $\mathcal{R}$ .

## 5 Experimental Setup

### 5.1 Construction of Experiment Datasets

We used two datasets covering the terrorism event of Australian Embassy bombing (AEB) in Jakarta, September 2004. They are the AEB and AEB-1000 datasets. Both datasets were created by downloading documents from an on-line news website and converting the documents to plain text. The seed words used for the extraction task are “*Australian Embassy*”, “*Australian Embassy Bombing*”, “*Suicide Bombers*” and “*Elisabeth Musu*”. Among them, Elisabeth Musu is a victim who was injured during the event. Based on the above seeds, other entity and relation instances about the event were determined by an expert familiar with the event as shown in Figure 1. In the figure, relation instances are



**Fig. 1.** Entity and Relation Instances of Australian Embassy Bombing Event

represented by directed edges. The numbers in brackets show the occurrences of the relation instance in all relevant documents.

**AEB** dataset has 100 documents consisting of 34 relevant documents and 66 irrelevant documents. The 34 relevant documents were selected to cover all instances of the bombing event. The 66 irrelevant documents were selected from more than 10,000 documents downloaded during the week after the event occurred. These documents were intentionally selected to describe other similar criminal events such as murdering and kidnapping. In other words, both the relevant and irrelevant documents describe some criminal events and they have a certain similarity content-wise. This also increases the level of difficulty in the document selection.

**AEB-1000** dataset has 1000 documents. The relevant documents in AEB-1000 are identical to that of AEB. In addition to the 66 irrelevant documents, we randomly selected 900 irrelevant documents for AEB-1000. With AEB-1000, we can evaluate the performance of our document selection strategies for a larger dataset.

There are altogether 7 entity classes and 10 relation classes that we are interested. The 7 entity classes are: Victim, Terrorist, Terrorist Organization(Org), Event, Location, Employer and Relative. And the 10 relation classes are:  $T_{Harm}$ ,  $O_{Harm}$ ,  $Connect\_to$ ,  $T_{Responsible\_for}$ ,  $O_{Responsible\_for}$ ,  $Member\_of$ ,  $Leader\_of$ ,  $Happen$ ,  $Work\_for$  and  $Relative\_of$ . Table 1 shows these 10 relation classes with their source and target entity classes. Table 2 shows more detailed information about the two datasets.



**Table 1.** 10 relation classes with their source and target entity classes

$\langle \text{Rel} \rangle$	$(\langle \text{SrcEnt} \rangle, \langle \text{TgtEnt} \rangle)$	$\langle \text{Rel} \rangle$	$(\langle \text{SrcEnt} \rangle, \langle \text{TgtEnt} \rangle)$
$T_{Harm}$	(Terrorist, Victim)	$O_{Harm}$	(Terrorist Org, Victim)
Connect_to	(Terrorist Org, Terrorist Org)	$T_{Responsible\_for}$	(Terrorist, Event)
$O_{Responsible\_for}$	(Terrorist Org, Event)	Member_of	(Terrorist, Terrorist Org)
Leader_of	(Terrorist, Terrorist Org)	Happen	(Event, Location)
Work_for	(Victim, Employer)	Relative_of	(Victim, Relative)

**Table 2.** Detailed Information of the two datasets

	$ \mathcal{E} $	$ \mathcal{R} $	# of Relevant docs	# of Optimal docs	Total docs
AEB	19	20	34	9	100
AEB-1000	19	20	34	9	1000

## 5.2 Construction of Extraction Patterns

The IE system chosen for the experiment is Crystal [2]. We have manually created a set of extraction patterns for extracting the entity and relation instances. These extraction patterns were created based on some common linguistic structures of the English language in order to be applied in a generic extraction task. For example:  $\langle \text{Subject} \rangle \langle \text{Verb} \rangle \langle \text{Object} \rangle$ ,  $\langle \text{Subject} \rangle \langle \text{Verb} \rangle \langle \text{Prepositional Phrase} \rangle$ ,  $\langle \text{Verb} \rangle \langle \text{Object} \rangle \langle \text{Prepositional Prase} \rangle$  and  $\langle \text{Subject} \rangle \langle \text{Prepositional Phrase} \rangle$  are four common structures we used. By constraining one or more part of each structure by words, we have the extraction patterns. For example:  $\langle \text{Subject} \rangle$  in one extraction pattern can be constrained by the Terrorist entity class, while the  $\langle \text{Object} \rangle$  is constrained by the Victim entity class. An extraction pattern is not going to fire on a sentence unless some constraints have been met. In other words, we use instances in the action and entity classes to guard the invocation of extraction patterns.

As we are interested in terrorism events, we use WordNet [3] to obtain some words for initializing the entity and action classes in the extraction patterns. These are generic words that can be used to describe the action classes relevant to terrorism and the names of already known terrorists. In our experiments, 21 terrorists’ names found on FBI website<sup>1</sup> and 54 terrorist organization’s names found on ICT website<sup>2</sup> have been included into entity class *Terrorist* and *Terrorist Organization* instance sets to form the extraction patterns.

## 5.3 Evaluation Settings

In our experiment, we set  $\alpha = 0.6$  to place a higher emphasis on the relation instances with respect to the seed entity instances. For CombineCompare, we

<sup>1</sup> <http://www.fbi.gov/mostwant/terrorists/fugitives.htm>

<sup>2</sup> <http://www.ict.org.il>

set  $\beta = 0.8$  to give more weight to documents containing larger number of relation instances already found. The experiment was conducted by running the extraction task for 45 iterations on AEB, and 50 iterations on AEB-1000.

We also propose a set of performance metrics defined below, which were evaluated after every 5 documents have been selected. These performance metrics focus on how much relevant instances the selected documents contain and how well each document selection strategy perform.

### 1. Evaluation on Extracted Entity and Relation Instances

Suppose we have all relevant entity instances in set  $\mathcal{E}_r$  and all relevant relation instances in set  $\mathcal{R}_r$ . To evaluate the resultant sets obtained in extraction task i.e.,  $\mathcal{E}$  and  $\mathcal{R}$ , let  $|\mathcal{E}_a|$  be the number of intersection between sets  $\mathcal{E}_r$  and  $\mathcal{E}$ , i.e.  $|\mathcal{E}_a| = |\mathcal{E}_r \cap \mathcal{E}|$ , and  $|\mathcal{R}_a|$  be the number of intersections between sets  $\mathcal{E}_r$  and  $\mathcal{R}$ , i.e.  $|\mathcal{R}_a| = |\mathcal{R}_r \cap \mathcal{R}|$ . The recall measure is defined as follows:

$$- \text{Recall}_{average} = \frac{1}{2}(\text{Recall}_{entity} + \text{Recall}_{relation})$$

where  $\text{Recall}_{entity} = \frac{|\mathcal{E}_a|}{|\mathcal{E}_r|}$  and  $\text{Recall}_{relation} = \frac{|\mathcal{R}_a|}{|\mathcal{R}_r|}$

### 2. Evaluation on Document Selection

Let  $L$  be the set of all relevant documents and  $S$  denote the set of selected documents. The precision and recall measures with respect to relevant documents are defined as follows:

$$- \text{Precision}_{rel\_doc} = \frac{|S \cap L|}{|S|}$$

$$- \text{Recall}_{rel\_doc} = \frac{|S \cap L|}{|L|}$$

Suppose there are  $v$  different optimal sets among all relevant documents (since the optimal set is usually not unique). Let  $\mathcal{O}$  denote the set of all optimal sets, i.e.,  $\mathcal{O} = \{O_1, O_2, \dots, O_v\}$ . We have  $|O_1| = |O_2| = \dots = |O_v|$ . The recall and precision measures respect to optimal set are defined as follows:

$$- \text{Recall}_{opt\_doc} = \frac{\max_{O_i \in \mathcal{O}} |O_i \cap S|}{|O_i|}$$

$$- \text{Precision}_{opt\_doc} = \frac{\max_{O_i \in \mathcal{O}} |O_i \cap S|}{|S|}$$

## 5.4 Ideal Document Selection Strategies

We introduce an ideal document selection strategy here to compare our proposed document selection strategies. The ideal selection strategy selects a document that gives the largest increase in the proportion of performance measurement during each iteration and the selected document must be a relevant document. We assume all documents have been manually annotated with entity and relation instances. Therefore, the score formula for ideal selection strategy is defined as follows:

$$score(d_i) = M(d_i)$$

where  $M$  refers to the improvement of a chosen performance metric brought by selecting document  $d_i$ .

Note that the ideal document selection strategy is not achievable in practice as we cannot accurately determine the instances in the documents to be selected. We however would like to use it to examine how far worst are the other document selection strategies.

## 6 Experimental Results

Figure 2(a) shows the  $Recall_{average}$  of AEB dataset. PartialMatch gives the best performance as it reaches almost perfect recall with the smallest number of iterations (documents selected). DiffCompare is the runner-up, followed by CombineCompare ( $\beta = 0.8$ ). PartialMatch extracted 95% entity and relation instances with almost less than half the number of documents compared to other pattern-based strategies.

We conclude from Figures 2(b) and 2(c) that PartialMatch consistently performs well in selecting relevant documents for the AEB dataset. It maintained perfect  $Precision_{rel.doc}$  until more than 20 iterations. While  $Precision_{rel.doc}$  of other strategies oscillate below 1 indicating that they are not able to select the relevant documents all the time. Although not shown in the figure, PartialMatch performs better on selecting the optimal documents. It selected 90% optimal documents at the 21th iteration, which is much better than DiffCompare (41th iteration) and CombineCompare (44th iteration).

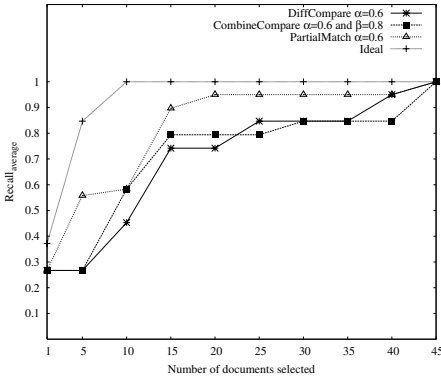
For the AEB-1000 dataset, PartialMatch reached 95%  $Recall_{average}$  in the 21th iteration as shown in Figure 2(d). This is followed by CombineCompare ( $\beta = 0.8$ ) and DiffCompare. The extraction task selected only 5% of the total number of documents and obtained almost all entity and relation instances with PartialMatch. In other words, even the worst pattern-based strategies can find more than 65% instances by selecting only 5% documents.

Although PartialMatch is the best among all strategies, it's performance is lower than that of Ideal selection. Figure 2(a) shows that the Ideal strategy reaches perfect  $Recall_{average}$  in the 9th iteration, while PartialMatch requires 44 iterations. Therefore, there is still some room for improvement.

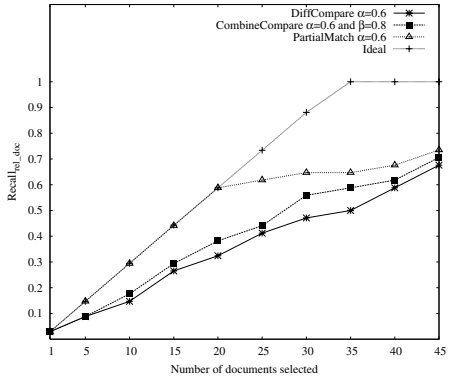
## 7 Conclusions

We have proposed a new event driven extraction task and four pattern-based document selection strategies in this paper. This task is applied to the terrorism event information extraction. Our objective is to select as few documents as possible to construct the event related entity and relation instances.

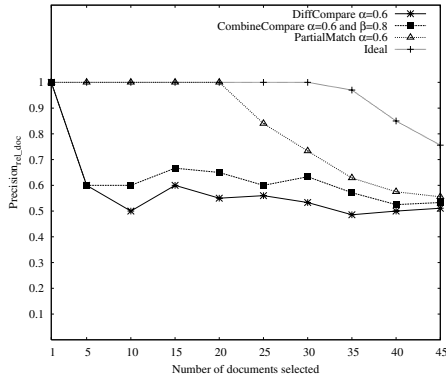
We have defined performance metrics to compare the proposed document selection strategies and conducted several experiments on 2 datasets. Experimental results conclude that our proposed strategies perform well on the extraction task.



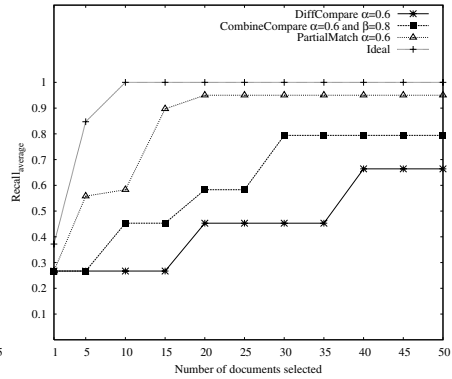
(a) Extracted Instances (AEB)



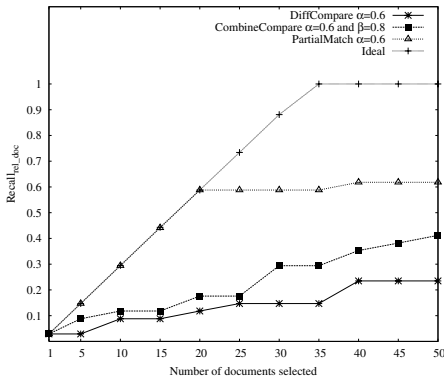
(b) Document selection recall (AEB)



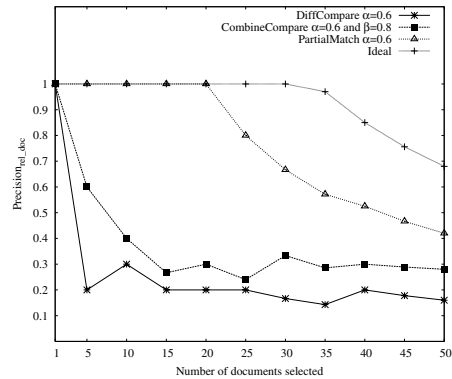
(c) Document selection precision (AEB)



(d) Extracted Instances (AEB-1000)



(e) Document selection recall(AEB-1000)



(f) Document selection precision(AEB-1000)

Fig. 2. Experimental Results

Among our proposed strategies, PartialMatch shows the best performance. Especially for a dataset containing 1000 documents, it managed to extract 95% of the required event related information by selecting only 5% of the documents.

## References

1. Finn, A., Kushmerick, N.: Active learning selection strategies for information extraction. In: Proceedings of ATEM. (2003)
2. Soderland, S., Fisher, D., Aseltine, J., Lehnert, W.: Crystal: Inducing a conceptual dictionary. In: Proceedings of the 14th IJCAI. (1995)
3. Fellbaum, C.: Wordnet: An electronic lexical database. MIT Press (1998)
4. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y.: Named entity recognition from diverse text types. In: Proceedings of Natural Language Processing 2001 Conference. (2001)
5. Huffman, S.: Learning information extraction patterns from examples. In: Proceedings of IJCAI-95 Workshop on new approaches to learning for natural language processing. (1995)
6. Riloff, E.: Automatically constructing a dictionary form information extraction tasks. In: Proceedings of the 11th National Conference on Artificial Intelligence. (1993)
7. Riloff, E.: Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th National Conference on Artificial Intelligence. (1996)
8. Riloff, E., Jones, R.: Learning dictionaries for information extraction by multi-level bootstrapping. In: Proceedings of the 16th National Conference on Artificial Intelligence. (1999)
9. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. (2002)
10. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of the Fifth ACM International Conference on Digital Libraries. (2000)
11. Allan, J., Papka, R., 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. (1998)
12. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th annual international conference on Research and development in information retrieval. (2004)
13. Wei, C.P., Lee, Y.H.: Event detection from online news documents for supporting environmental scanning. *Decis. Support Syst.* **36** (2004) 385–401
14. C, Michael., J, Xu., Chen, Hsinchun.: Extracting Meaningful Entities from Police Narrative Reports. In: Proceedings of the National Conference for Digital Government Research. (2002)