Investigations on Event Evolution in TDT

Proceedings of HLT-NAACL 2003
Student Research Workshop , pp. 43-48
Edmonton, May-June 2003

Juha Makkonen
Department of Computer Science,
P.O. Box 26, 00014 University of Helsinki
Finland
jamakkon@cs.helsinki.fi

Abstract

Topic detection and tracking approaches monitor broadcast news in order to spot new, previously unreported events and to track the development of the previously spotted ones. The dynamical nature of the events makes the use of state-of-the-art methods difficult. We present a new topic definition that has potential to model evolving events. We also discuss incorporating ontologies into the similarity measures of the topics, and illustrate a dynamic hierarchy that decreases the exhaustive computation performed in the TDT process. This is mainly work-in-progress.

1 Introduction

A fairly novel area of retrieval called *topic detection and tracking* (*TDT*) attempts to design methods to automatically (1) *spot new*, *previously unreported events*, and (2) *follow the progress of the previously spotted events* (Allan et al., 1998c; Yang et al., 1998).

Our contribution deals with three problems in TDT. Firstly, we present a new definition for a topic that would model the *event evolution*, i.e., the changing nature of a topic. The previous event definitions do not really lend themselves to this change. Secondly, we investigate an approach suggested by Makkonen, Ahonen-Myka and Salmenkivi (2002). They partitioned the termspace into four semantic classes and represented each class with a designated vector. Unlike the term-weighting model of Yang *et al.* (2002) this approach enables the use of introduction of different similarity measures for each semantic class. We formalize the comparison method and suggest a kNN approach based on this formalization. Thirdly, we suggest the use of dynamic hierarchies in a TDT system that would decrease the exhaustive computation of the first story detection. In practice this means that we import text categorization on top of TDT. The purpose of this paper is to outline the main aspects of our ongoing and future work. As this is mainly work-in-progress, we do not have empirical motivation for our work.

This paper is organized as follows: We will discuss the problems of TDT in Section 2 In Section 3 we examine the definitions of an event and a topic. Section 4 presents a novel event representation and an approach to measure the similarity of such elements. In Section 5 we deal with dynamic hierarchies. In Section 6 we discuss our conclusions.

2 Problems in TDT

The events are taking place in the world, and some of them are reported in the news. A TDT system does not perceive the events themselves, rather makes an effort in deducing them from the continuous news-stream – which is in a sense like the shadows on the wall in Plato's cave analogy. Given this setting, what is it that we are trying to model?

Typically, the text categorization is conducted using some machine learning system (Sebastiani, 2002; Yang and Liu, 1999). Such system is taught to recognize the difference between two or more predefined classes or categories by providing a good number of pre-labeled samples to learn from. As to classes and word frequencies, this training material is assumed to lend itself to the same underlying distribution as the material that is to be categorized. More formally, the documents $X = \{x_1, x_2, \ldots, x_{|X|}\}$ and their labels $C = \{c_1, c_2, \ldots, c_{|C|}\}$ yield to a unknown distribution. This distribution is expressed as a function $\check{h}$ that assigns to each document-label pair

$$\{\langle x_i, c_j \rangle \in X \times C|\ 1 \leq i \leq |X|, 1 \leq j \leq |C|\}$$

a boolean value indicating their relevance, i.e., $\check{h} : X \times C \rightarrow \{-1, 1\}$. The task of classification is to come up with a hypothesis $h : X \times C \rightarrow \{-1, 1\}$ that represents

$\check{h}$, practically, with the 'highest' accuracy. This accuracy is evaluated with a pre-labeled testing material.

Now, with TDT the problem is different. Let us assume that the documents and events yield to an unknown distribution represented by the function $\check{g} : X \times E \rightarrow \{-1, 1\}$ that assigns each document $x_i \in X$ a boolean value indicating whether it discussed event $e_j \in E$ or not. The problem is that domain of $E = \{e_1, e_2, \ldots, e_{|E|}\}$ is time-dependent. The hypothesis $g : X \times E \rightarrow \{-1, 1\}$ built from the training data does not work with evaluation data, because these two data sets do not discuss the same events. Moreover, the events are very small in size compared to categories and their identity, that is, the most important terms evolve over time. We can, however, model *similarity* between two documents. By examining the pair-wise comparisons in the training set, we can formulate a hypothesis $k : X \times X \rightarrow \{-1, 1\}$ that assigns the pair $\langle x_i, x_j \rangle \in X \times X$ a boolean value 1 if the documents discuss the same event, -1 otherwise. Any two documents of same event are (ideally) *similar in a similar way*. This somewhat trivial observation has some implications worth mentioning.

Firstly, by definition news documents report changes, something new with respect to what is already known. This would lead to think that the identity of an event eludes all static representations and that the representation for a topic would have to adapt automatically to the various changes in the reporting of the event.

Secondly, so far the parameters and thresholds of the state-of-the-art methods in IR have tried to capture this similarity of similarity, but there does not seem be a representation expressive enough (Allan et al., 2000).

Thirdly, the detection and tracking is based on pair-wise comparisons which requires exhaustive computation. Yang *et al.* (2002) suggested topic-categories that could be used to limit the search space of the first-story detection. However, building topic-categories automatically is difficult. In the following we outline some suggestions to these problems: event modeling, event representation and decreasing the computational cost.

3 Events and Topics

Although the concept of *event* might seem intuitively clear and self-explanatory, formulating a sound definition appears to be difficult. Predating TDT research, numerous historians and researchers of political science have wrestled with the definitions (Falk, 1989; Gerner et al., 1994). What seems to be somewhat agreed upon is that an event is some sort of activity conducted by some agent and taking place somewhere at some time.

Definition 1 *An event is something that happens at some specific time and place (Yang et al., 1999).*

This initial definition was adopted to TDT project and it is intuitively quite sound. Practically all of the events of the TDT test set yield to the temporal proximity ("burstiness") and the compactness. However, there is also a number of problematic cases which this definition seems to neglect: events which either have a long-lasting nature (Intifada, Kosovo–Macedonia, struggle in Columbia), escalate to several large-scale threads or campaigns (September 11), or are not tightly spatio-temporally constrained (BSE-epidemics).

The events in the world are not as autonomous as this definition assumes. They are often interrelated and do not necessarily decay within weeks or a few months. Some of these problematic events would classify as *activities* (Papka, 1999), but when encountering a piece of news, we do not know *a priori* whether it is a short term event or long term activity, a start for a complex chain of events or just a simple incident.

Definition 2 *An event is a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences (Cieri et al., 2002).*

This is basically a variant of Definition 1 that in some sense tries to address the autonomy assumption. Yet, it opens a number of questions as to what are the necessary preconditions for certain event, an oil crisis, for example. What are the necessary preconditions and unavoidable consequences of Secretary Powell's visit to Middle East or of suicide-bombing in Ramallah?

Definition 3 *A topic is an event or an activity, along with all related events and activities (Cieri et al., 2002).*

Here, Cieri *et al.* explicate the connection between a topic and an event: they are more or less synonyms. *Rules of interpretation* that have been issued to help to draw the line and to attain consistency. In TDT, there are eleven topic types that tell what kind of other topic types are relevant. The topic type of the topic is determined by the *seminal event*. Since TDT2 and TDT3 corpora are produced along this guideline, this is in a sense the *de facto* definition.

Definition 4 *A topic is a series of events, a narrative that evolves and may fork into several distinct topics.*

Definition 4 makes an attempt at addressing the changing or evolving nature of a topic. A seminal event can lead to several things at the same time and the connection between the various outcomes and the initial cause become less and less obvious as the events progress. As a practical consequence, the *event evolution* (Yang et al., 1999; Papka, 1999) causes changes in the vocabulary, especially in the crucial, identifying terms.

The news documents are temporally linearly ordered, and the news stories can be said to form series of different lengths. Identifying these chains as topics is motivated by Falk's investigations on historical events (Falk, 1989). A narrative begins as soon as the first story is encountered. Then the narrative is developed into one or more directions: simple events, like plane accidents might not have as many sub-plots as a political scandal, a war or economical crises. Then, at some point one could say the latest story is so different from the initial one that it is considered a first story for a new event. However, there could remain some sort of link that these two topics (narratives) are somehow relevant. Hence, this kind of a narrative has a beginning, a middle and an end. An event evolution is illustrated in Figure 1.

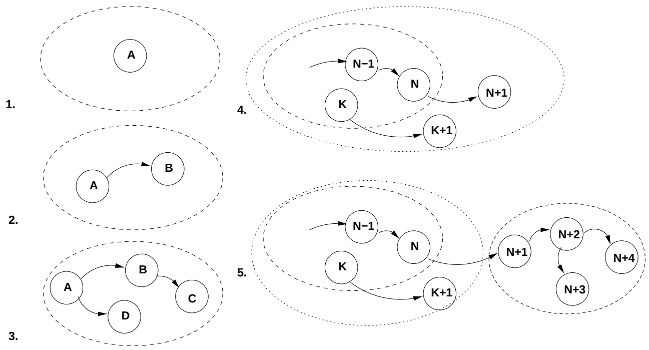

Figure 1: An example of event evolution.

Initially, in phase 1 we have only one document, a first story A, an it constitutes an event that is depicted by the dashed line. Then in phase 2, document B is found relevant to this event. Since it is found similar to A, there is link in between them. In phase 3 there are two more relevant documents: C and D. The former is more found similar to B than to A, and thus it continues the off-spring started by B. On the contrary, D appears closer to A and thus it starts a new direction. Phase 4 shows two stories, $N+1$ and $K+1$ outside the dashed ellipse. This represents a situation, where the vocabulary of the two expulsed documents is diverging from the rest of the documents, i.e., the *inner cohesion* of the topic is violated too much. The dotted ellipse represents the domain of possible *topical shifts*, i.e., stories that lead too far from the original topic. They are still regarded as part of the topic, but are on the brink of diverging from the topic and hence candidates for new first stories or seminal events. Finally, in phase 5 the separation takes place: Three new documents, $N+2$, $N+3$ and $N+4$, are found similar to $N+1$. As a result, document $N+1$ is separated into its own topic. Note that there is no follow-ups for $K+1$, and therefore it is not cut off.

The problem of text summarization is similar to detecting topical shifts: traces of all the main topics occurring in the given text need to be retained in the summarization. On the other hand, text segmentation shares some qualities with the topic shift detection. *Lexical cohesion* (Boguraev and Neff, 2000) has been employed in the task as well as in text segmentation (Stokes et al., 2002).

A model of Definition 4 has many open issues. For example, what is the topic representation and what kind of impact will there be on the evaluation? We will try to address the former question in the following.

4 Multi-vector Event Model

It has been difficult to detect two distinct train accidents or bombings as different events (Allan et al., 1998a). The terms occurring in the two documents are so similar that the term-space or the weighting-scheme in use fails to represent the required very delicate distinction. Furthermore, Allan, Lavrenko and Papka suspect that only a small number of terms is adequate to make the distinction between different news events (Allan et al., 1998b). Intuitively, when reporting two different train accidents, it would seem that the location and the time, possibly some names of people, are the terms that make up the difference. Papka observes that when increasing the weights of noun phrases and dates the classification accuracy improves and when decreasing them, the accuracy declines (Papka, 1999).

4.1 Event Vector

A news document reporting an event states at the very barest *what* happened, *where* it happened, *when* it happened, and *who* was involved. The automatic extraction of these facts for natural language understanding is quite troublesome and time-consuming, and could still perform poorly. Previous detection and tracking approaches have tried to encapsulate these facts in a single vector. In order to attain the delicate distinctions mentioned above, to avoid the problems with the term-space maintenance and still maintain robustness, we assign each of the questions a *semantic class*, i.e., i.e. groups of semantically related words, similarly to approach suggest by Makkonen *et al.* (2002). The semantic class of LOCATIONS contains all the places mentioned in the document, and thus gives an idea, where the event took place. Similarly, TEMPORALS, i.e., the temporal expressions name an object, that is, a point or an interval of time, and bind the document onto the time-axis. NAMES are proper noun phrases that represent the people or organizations involved in the news story. What happened is represented by 'normal' words which we call TERMS.

This approach has an impact on the document and the event representations. Instead of having just one vector, we issue four sub-vectors – one for each semantic class as illustrated in Figure 2.

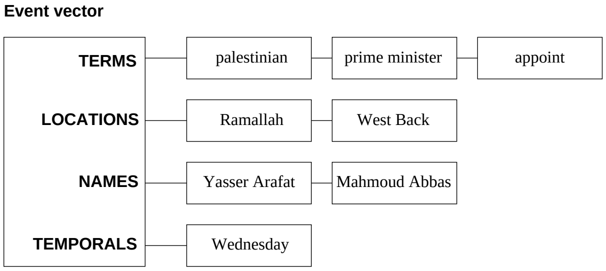

Figure 2: *"RAMALLAH, West Bank - Palestinian leader Yasser Arafat appointed his longtime deputy Mahmoud Abbas as prime minister Wednesday, . . . "* (AP: Wednesday, March 19, 2003)

4.2 Similarity of Hands

One could claim that the meaning of a word is in the word's relation to other words without getting too deep into philosophical discussions as to what and how the meaning is. This meaning, that is, relation, can be represented in an ontology, where similar terms relate to each other in different manner than dissimilar ones.

The similarity of event vectors is determined *class-wise*: Each semantic class has its own similarity measure, and the over-all similarity could be the weighted sum of these measures, for example. The interesting thing is that now we can introduce semantics into the vector-based similarity by mapping the terms of a semantic class onto a formal space. Each pair of terms in this space has a similarity, i.e., a distance. Two TEMPORAL terms relate to each other on the time-axis, and the similarity of two LOCATION terms can be based on a geographical proximity represented in an ontology. For example, the utterances *next week* and *the last week of March 2003* do not coincide on the surface, but when evaluated with respect to the utterance time, the expressions refer to the same temporal interval. Similarly, *London* and *Thames* can be found relevant based on an spatial ontology. Similarity in these ontologies could be a distance on the time-axis or a distance in a tree, as we have previously noted (Makkonen et al., 2003).

Now, let us present the above discussion more formally. Each term in the document is a member of exactly one semantic class. Hence, the documents are composed of the union of semantic classes, or equivalently, the document is a structure of a language specified by the unary relations that represent the semantic classes.

Definition 5 *Let W be a universe and let Σ be a language consisting of m unary relations $\Sigma = \{S_1, S_2, \ldots, S_m \mid S_i \subseteq W\}$. A model is Σ-structure $\mathcal{A} = (W, \Sigma)$.*

Now, consider W as the set of natural language terms and Σ as the set of semantic classes. A document representation would be a Σ-structure consisting of terms

$$\{w \in W \mid\ w \in \bigcup_{i=1}^{m} S_i\},$$

i.e., a document is simply a union of the semantic classes. The class-wise similarity of two such structures would be as follows:

Definition 6 *Let δ_i be a function $\delta_i : W \times W \to I\!R$ that indicates the similarity of two elements in S_i The similarity of two Σ-structures is a function $sim : W \times W \to I\!R^m$ such that*

$$sim(\mathcal{A}, \mathcal{B}) = \delta_i(S_i^{\mathcal{A}}, S_i^{\mathcal{B}})_{i=1}^{m} = v_i. \tag{1}$$

This type of similarity we call the similarity of hands [1].

Hence, the similarity of two documents, $\mathcal{A}$ and $\mathcal{B}$, would be a vector $(v_1, v_2, \ldots, v_m) \in I\!R^m$. There are many ways to go about turning the vector into a single score (Makkonen et al., 2002). One way is to define the similarity as a weighted sum of each value of δ_i, i.e.,

$$sim(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{m} \alpha_i\ \delta_i(S_i^{\mathcal{A}}, S_i^{\mathcal{B}}), \tag{2}$$

where $\alpha_i \in I\!R$ is the relative weight of class S_i. The similarities $\delta_i(S_i^{\mathcal{A}}, S_i^{\mathcal{B}})$ have also been interpreted as van Rijsbergen's (van Rijsbergen, 1980) similarity coefficients.

Unlike detective stories, news documents give away the plot in the first few sentences. Therefore, the similarity measure could exploit the ranking, the ordinal of the sentence in which the term appears, in weighting the class-wise similarity.The *rank-score* of a term t occurring m times is

$$rs(t) = \sum_{k=1}^{m} \frac{1}{\log{(t_k)}}, \tag{3}$$

where t_k is the ranking of the kth instance of term t. Hence the similarity δ_i would yield

$$\delta_i^{ranks}(S_i^{\mathcal{A}}, S_i^{\mathcal{B}}) = \sum_{k=1}^{n} rs(t_k), \tag{4}$$

where term t_k occurs n times in intersection $S_i^{\mathcal{A}} \cap S_i^{\mathcal{B}}$.

Currently we are experimenting with similarity of hands technique as a *relevance score* (Yang et al., 2000) for ranking the k nearest neighbours for each semantic

[1] Consider a simple game where one would have to determine the similarity of two hands of cards of arbitrary size (up to 52) drawn from two distinct decks and assume that there is a designated similarity measure for each suit. For example, with hearts low cards could be of more value. Furthermore, the suits could be weighted, i.e., clubs could be trump and unchallenged clubs would lead to dissimilarity.

class. In other words, we find the k nearest events with respect to TEMPORAL, k nearest events with respect to NAMES, etc. In a sense, each semantic class votes for k candidates based on the relevance score and the respective weight of the semantic class. Once we have the four sets of candidates, we elect the one with highest number of votes.

Hence, let $\mathcal{D} = \{D_1, D_2, \ldots, D_l\}$ be the set of previous Σ-structures (i.e., events). The function $votes : W \times \mathbb{N} \times W^l \rightarrow W^{km}$ returns m vectors of Σ-structures of length k consisting of structures closest to $\mathcal{A}$ with respect to relation S_i. In other words,

$$\begin{aligned} votes(\mathcal{A}, k, \mathcal{D}) &= (\delta_i(\mathcal{A}, D_k)_{i=1}^m)_{k=1}^l \\ &= ((X_1)_1^k, (X_2)_1^k, \ldots, (X_m)_1^k), \end{aligned}$$

where $(X_i)_1^k$ is a length-k vector of Σ-structures closest to $\mathcal{A}$ with respect to relation S_i. The election is a function $elect : W^{km} \rightarrow W$ such that

$$elect((X_1)_1^k, (X_2)_1^k, \ldots, (X_m)_1^k) = \bigcap_{p=1}^{k} \bigcap_{i=1}^{m} (X_i)p.$$

Quite obviously, the intersection is too strong a function in this case. Some vector $(X_i)_1^k$ might be empty which would make the intersection empty as well. However, we believe that it would be easier to find optimal weights for the semantic classes via this voting scheme than trying to optimize Equation 2, because there are less parameters.

5 Dynamic Hierarchies

One of the problems that plagues many TDT efforts is the need to compare each incoming document with all the preceding documents. Even if we issue a time-window and have a straight-forward similarity measure, the number of required comparisons increases drastically as new documents come in. There have been efforts to decrease the amount of work by centroid vectors (Yang et al., 2000), and by building an *ad hoc* classifier for each topic-category (Yang et al., 2002), for example.

We suggest the we adopt text categorization on top of topic detection and tracking, similar to Figure 3. There has been good results in text categorization (see, e.g., (Yang and Liu, 1999; Sebastiani, 2002)) The predefined categories would form the static hierarchy – the IPTC Subject Reference System [2], for example – on top of all event-based information organization, and the models for the categories could be built on the basis of the test set.

Below the static hierarchy there would be a dynamic hierarchy that evolves as new documents come in and new topics are detected. There is also a time-window to limit the temporal scope. Once a topic expires, it is removed from the dynamic hierarchy and archived to a news repository of lower operational priority.

The use of static hierarchy has some of the benefits the topic-categories of Yang *et al.* (2002) had. It decreases the search space and enables a category-specific weighting-scheme for terms. For example, when a document is categorized to the class 'science', there is no need to compare it against the events of any other class; ideally, all the relevant events have also been categorized to the same class.

6 Conclusion

We have discussed three problems relating to TDT and mainly its event evolution. The novel topic definition allows the topic to evolve into several directions and ultimately to distinguish new topics. A semantic class-based event vector enables harnessing of domain specific ontologies, such as the time-axis and the geographic distances. Finally, we presented a TDT system with dynamic hierarchies that would cut down the excessive computation required in the TDT process.

Our previous results were done with a Finnish online news corpus smaller than the TDT corpora (Makkonen et al., 2002). The use of semantic classes proved to be beneficial. We have also built a temporal expression scheme and a geographical ontology for TDT purposes (Makkonen et al., 2003). In this paper, all our discussions were preliminary and should be regarded as such. In the future we will work to motivate these mostly intuitive theories with empirical results.

References

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998a. Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, February.

James Allan, Victor Lavrenko, and Ron Papka. 1998b. Event tracking. Technical Report IR – 128, Department of Computer Science, University of Massachusetts.

James Allan, Ron Papka, and Victor Lavrenko. 1998c. On-line new event detection and tracking. In *Proc. ACM SIGIR*, pages 37–45.

James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *Proc. 9th Conference on Information Knowledge Management CIKM*, pages 374–381, McClean, VA USA.

[2] International Press Telecommunications Council, http://www.iptc.org

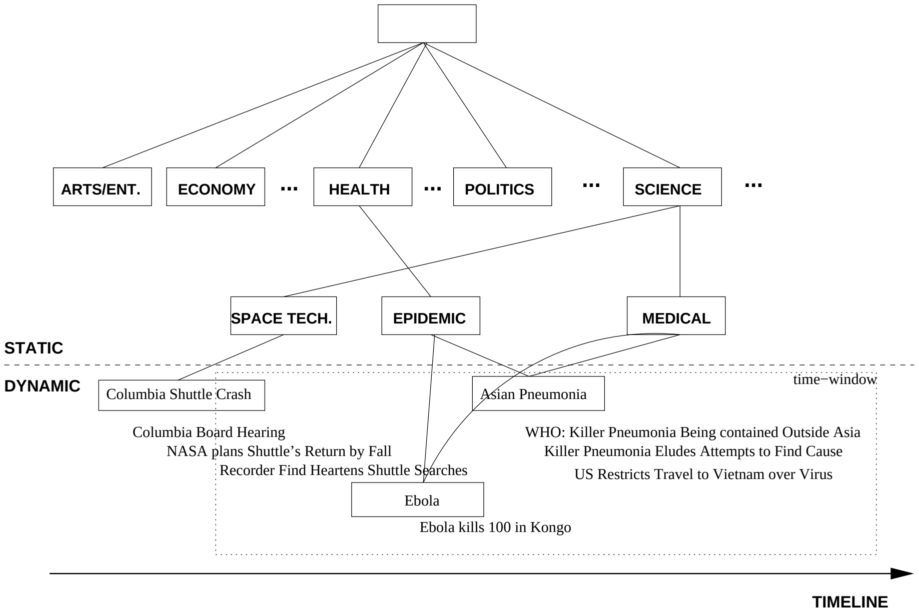
ARTS/ENT.
ECONOMY
...
HEALTH
...
POLITICS
...
SCIENCE
...
SPACE TECH.
EPIDEMIC
MEDICAL
STATIC
DYNAMIC
Columbia Shuttle Crash
Asian Pneumonia
time−window
Columbia Board Hearing
NASA plans Shuttle's Return by Fall
Recorder Find Heartens Shuttle Searches
WHO: Killer Pneumonia Being contained Outside Asia
Killer Pneumonia Eludes Attempts to Find Cause
US Restricts Travel to Vietnam over Virus
Ebola
Ebola kills 100 in Kongo
TIMELINE

Figure 3: A dynamic hierarchy with a static IPTC taxonomy on top and a topic-based time-varying structure on the bottom.

Branimir K. Boguraev and Mary S. Neff. 2000. Lexical cohesion, discourse segmentation and document summarization. In *Proc.RIAO'2000 (Recherche d'Informations Assistee par Ordinateur)*, pages 237–246, Paris.

Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman. 2002. Corpora for topic detection and tracking. In James Allan, editor, *Topic Detection and Tracking – Event-based Information Organization*, chapter 3, pages 33–66. Kluwer Academic Publisher.

Pasi Falk. 1989. The past to come. *Economy and Society*, 17(3):374–394.

Deborah J. Gerner, Philip A. Schrodt, Ronald Francisco, and Julie L. Weddle. 1994. The analysis of political events using machine coded data. *International Studies Quarterly*, 38:91–119.

Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. 2002. Applying semantic classes in event detection and tracking. In *Proc. International Conference on Natural Language Processing (ICON'02)*, Mumbai, India.

Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. 2003. Topic detection and tracking with spatio-temporal evidence. Accepted in ECIR 2003.

Ron Papka. 1999. *On-line New Event Detection, Clustering and Tracking*. Ph.D. thesis, Department of Computer Science, University of Massachusetts.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. 2002. Segmenting broadcast news streams using lexical chains. In *Proc. STarting AI Researchers Symposium, (STAIRS 2002)*, pages 145–154, Lyon.

C. J. van Rijsbergen. 1980. *Information Retrieval*. Butterworths, 2nd edition.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. ACM SIGIR*, pages 42–49, Berkley.

Yiming Yang, Thomas Pierce, and Jaime Carbonell. 1998. A study on retrospective and on-line event detection. In *Proc. ACM SIGIR*, pages 28–36, Melbourne.

Yiming Yang, Jaime Carbonell, Ralf Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. 1999. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval*, 14(4):32 – 43.

Yiming Yang, Thomas Ault, Thomas Pierce, and Charles Lattimer. 2000. Improving text categorization methods for event detection. In *Proc. ACM SIGIR*, pages 65–72.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proc. ACM SIGKDD (to appear)*, Edmonton, Canada.