New Event Detection and Topic Tracking in Turkish

Fazli Can, Seyit Kocberber, Ozgur Baglioglu, Suleyman Kardas, H. Cagdas Ocalan, and Erkan Uyar
Bilkent Information Retrieval Group, Computer Engineering Department, Bilkent University, Ankara, Turkey 06800. E-mail: {canf,ozgurb,skardas,hocalan,euyar}@cs.bilkent.edu.tr, seyit@bilkent.edu.tr

Topic detection and tracking (TDT) applications aim to organize the temporally ordered stories of a news stream according to the events. Two major problems in TDT are new event detection (NED) and topic tracking (TT). These problems focus on finding the first stories of new events and identifying all subsequent stories on a certain topic defined by a small number of sample stories. In this work, we introduce the first large-scale TDT test collection for Turkish, and investigate the NED and TT problems in this language. We present our test-collection-construction approach, which is inspired by the TDT research initiative. We show that in TDT for Turkish with some similarity measures, a simple word truncation stemming method can compete with a lemmatizer-based stemming approach. Our findings show that contrary to our earlier observations on Turkish information retrieval, in NED word stopping has an impact on effectiveness. We demonstrate that the confidence scores of two different similarity measures can be combined in a straightforward manner for higher effectiveness. The influence of several similarity measures on effectiveness also is investigated. We show that it is possible to deploy TT applications in Turkish that can be used in operational settings.

Introduction

Information explosion has new dimensions with the advances in information technologies. For example, the number of news resources on the World Wide Web has exponentially increased in the last decade. Multiresource news portals, a relatively new development, receive and gather news from several Web news providers. More advanced versions of these portals aim to make news stories more accessible by providing event-based information organization. Topic detection and tracking (TDT) applications aim to organize the temporally ordered stories of a news stream. Such event-based organizations facilitate an abstraction and aim to prevent overwhelming news consumers, which can be caused by too many unconnected stories (Hereafter, the words "news," "story," and "document" as well as "effectiveness" and "performance" are used interchangeably.)

Services for current events are popular on the Web. Commercial news portal examples with such services include Google News (http://news.google.com/) and NewsIsFree (http://www.newsisfree.com/). Research-oriented examples include NewsInEssence (Radev, Otterbacher, Winkel, & Balir-Goldensohn, 2005), NewsBlaster (McKeown et al., 2002), and multilanguage news services developed by the Europeans Commission's Joint Research Center (Pouliquen, Steinberger, Ignat, Kasper, & Temnikova, 2004).

In TDT, an event is defined as something that happens at a given "place and time, along with all the necessary preconditions and unavoidable consequences" (TDT, 2004, p. 4). For example, an event might be a car accident or a meeting. In TDT studies, a topic is defined as "*a seminal event* or *activity* with all directly related events and activities" (TDT, 2004, p. 4). In this context, an activity is defined as a connected series of events that have a common focus or purpose. Accordingly, a TDT activity may be a disaster relief effort, an election campaign, or an investigation. Note that the concept of *topic* in TDT is different from the notion of *topic* in normal discourse. One might normally think of a topic as something broad such as "accidents;" however, a TDT topic is limited to a specific accident (TDT, 2004).

The most influential research effort in this area is the TDT research initiative. In this work, we study two of the five tasks that are defined by this initiative. They are:

- *New Event Detection* (NED): aims to recognize the first story for a new event that has not been discussed before. This problem also is referred to as *first story detection*.
- *Topic Tracking* (TT): aims to find all other stories on a topic in the stream of arriving stories. In TT, the system is provided with a small number of stories (usually 1–4) known to be on the same topic.

The other TDT tasks are Story Segmentation, Topic Detection (Cluster Detection), and Story Link Detection.

Contributions

In this study, we

- Present a search-based, language-independent TDT test-collection-construction method implemented as a system called *ETracker*. It is inspired by a parallel method used in

Received August 4, 2008; revised October 1, 2009; accepted October 1, 2009

© 2010 ASIS&T • Published online 12 January 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21264

the TDT research initiative (Cieri, Strassel, Graff, Martey, Rennert, & Liberman, 2002; TDT, 2004).

- Present the characteristics of a large-scale TDT test collection, BilCol-2005 (Bilkent TDT Collection for the Year 2005), which was constructed using ETracker. It contains 209,305 news stories and 80 annotated events. BilCol-2005 is available to other researchers as the first test collection prepared for TDT studies in Turkish.
- Investigate the NED and TT problems in Turkish and provide pioneering benchmark observations.
- Show that different similarity measures can be used together for improving NED and TT effectiveness.
- Supply practical recommendations for the implementation of TDT systems in Turkish.

The rest of the article is organized as follows. First, we present a review of related studies. This is followed by our TDT test-collection-construction method, the characteristics of the constructed test collection, and a description of the evaluation methodology. Then, we present our NED and TT methods, the experimental environment, and the experimental results. We conclude the article with a summary of findings, some recommendations for the implementations of TDT applications in Turkish, and future research pointers.

Related Work

The new event-detection problem has not been studied prior to the TDT research initiative (Papka, 1999, p. 29). It was sponsored by NIST and continued between 1997 and 2004 (TDT, 2008). The compilation edited by Allan (2002a) is an excellent resource on this research initiative; it covers issues such as TDT evaluation, probabilistic and cluster-based approaches, statistical models, translingual topic tracking, and more.

In one commonly used method to solve the NED problem, the newest story is compared with the earlier stories to decide if it is different (i.e., dissimilar). "Unique" (i.e., different enough) stories are treated as the first stories of new events. The origins of this approach can be seen in IR; namely, in single-pass incremental document clustering (van Rijsbergen, 1979, p. 52) or in general cluster analysis literature (Anderberg, 1973, chap. 7). In practice, the use of such an approach is inefficient or unfeasible without resorting to employing a considerable amount of hardware resources (Luo, Tang, & Yu, 2007). A solution to this efficiency problem is the sliding time-window concept (see Figure 1). In the methods based on this concept, a new story is compared with only the members of a time window that contains the most recent predefined number of stories (Papka, 1999; Yang, Pierce, & Carbonell, 1998). Here, the assumption is that the stories related to an event are near to each other in terms of their occurrence in time. In this study, we also use the sliding time-window concept for NED.

Clustering concepts are used in various TDT studies. Yang, Pierce, and Carbonell (1998) used hierarchical and nonhierarchical document-clustering algorithms to solve the NED problem. In their approach, they paid attention to temporal information and used a time-decay function, making older documents have less influence on later decisions. As a part of the initial TDT research initiative, in his dissertation, Papka (1999) employed link-based clustering concepts in various TDT tasks. Allan, Lavrenko, and Jin (2000) studied the difficulties of finding new events with the traditional single-pass clustering approach, and showed that with certain assumptions, NED effectiveness can be predicted from that of TT since these two tasks are related. In this article, our NED method, which is based on the time-window concept, is inspired by the incremental clustering strategy and is similar to the ones available in the literature.

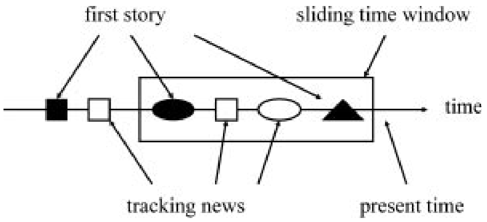


FIG. 1. Sliding time-window (different shapes represent different events).

Combining the results of different approaches has been studied by various research groups. Hatzivassiloglou, Gravano, and Maganti (2000) studied the problem of combining the results of different similarity functions and proposed a theoretically justified statistical model that performs as good as or better than that of an exhaustive search. Stokes and Carthy (2001) used a composite document representation that involves concept representation based on lexical chains. Yang et al. (2002) studied a combination system called *BORG* (Best Overall Results Generator for tracking) by using the results of various classifiers and examining their decision error tradeoff (DET) curves. Kumaran and Allan (2004) examined the effects of stop words and the combination of different document vectors (named entity vectors, nonnamed entity vectors) on NED. In this work, we examine the effects of using named entities, word stopping, and combining the results of different similarity measures on NED in Turkish.

TDT-based applications are becoming especially crucial in a new age which is prone to terrorist attacks (He et al., 2008). Luo et al. (2007) studied a practical new-event-detection system using IBM's Stream Processing Core middleware (Amini, 2007). Allan et al. (2005) presented the deployment of their TDT clustering technology in practical settings and the problems they faced when they take topic detection from evaluation to practice. We present our initial experience in practical settings in Can et al. (2008a) and in Kardaş (2009).

There is limited research on TDT in Turkish. This is partly due to the fact that there is no standard TDT test collection for this language since the preparation of such collections involves a significant amount of work. To the best of our knowledge, Kurt (2001) conducted the only TDT study for Turkish other than ours using 46,530 stories and 15 annotated events. Our communication with Kurt revealed that his test collection has been misplaced and is unavailable for further research.

TABLE 1. Information about distribution of stories among news sources in BilCol2005.

News source	No. of news stories	Percent of all stories	Download amount (MB)	Net amount (MB)	Average no. of words per story
CNN Türk	23,644	11.3	1,008.3	66.8	271
Haber 7	51,908	24.8	3,629.5	107.9	238
Milliyet Gazetesi	72,233	34.5	508.3	122.5	218
TRT	18,990	9.1	937.9	18.3	121
Zaman Gazetesi	42,530	20.3	45.3	33.7	97
Total	209,305	100.0	6,129.3	349.2	196*

*Different from the weighted sum of the average word lengths due to rounding error.

TDT Test-Collection Construction and BilCol-2005

In TDT, a test collection contains several news articles in temporal order. Among these articles, the first stories corresponding to a set of new seminal events and their tracking stories are identified by human annotators. In this section, we describe a search-based language-independent TDT test collection construction method inspired by a parallel approach used in the TDT research initiative (TDT, 2004). It is implemented in the form of a topic annotation system called *ETracker* (Öcalan, 2009).

For the construction of the news story collection, we used five different Turkish news Web sources: CNN Türk (http://www.cnnturk.com), Haber 7 (http://www.haber7.com), *Milliyet Gazetesi* (http://www.milliyet.com.tr), TRT (http://www.trt.net.tr), and *Zaman Gazetesi* (http://www.zaman.com.tr). From these sources, we downloaded all articles of the Year 2005 that have a timestamp in terms of day, hour, and minute. Duplicate or near-duplicate documents of this initial collection are eliminated by using a simple method: Stories with the same timestamp coming from the same source and with identical initial three words are assumed as duplicate or near-duplicate. We eliminated about 16,000 stories by this method. Such duplicates were caused by interrupted crawling or multiple identical postings of the news providers. The size of our test collection is comparable to those of the TDT research initiative (TDT, 2004). More information about our corpus, BilCol-2005, is provided in Table 1.

Topic Profiles

For each topic, the selected annotators are required to complete a topic profile (TDT, 2004). This process aims to provide documentation for the topics being annotated and helps annotators to properly investigate the related seminal event. The profiles also are used during the annotation process (discussed later). A topic profile has the following elements.

- Topic title: a brief phrase which is easy to recall and reminds the topic;
- Event summary: a summary of the seminal event with one or two sentences;
- What: what happened during the seminal event;
- Who: who was involved (people, organization etc.) during the seminal event;
- When: when the seminal event occurred;
- Where: where the seminal event happened;
- Topic size: annotator's estimate of topic size;
- Seed: the first story about the seminal event (the document number of the story in the collection); and
- Topic type: defined later.

Like the ones used in the TDT research initiative (Cieri et al., 2002; TDT, 2004), there are 13 topic types: elections, scandals/hearings, legal/criminal cases, natural disasters, accidents, acts of violence or war, science and discovery news, financial news, new laws, sports news, political and diplomatic meetings, celebrity/human interest news, and miscellaneous news.

Topic Annotation: ETracker System

Annotators selected their own topics; like that of TDT (2004), no effort was made to nurse an equal representation of each news source or month in the final set of selected events/topics. At the same time, by using the topic profiles, we watched the selected topic types to make sure that we have a wide variety of topic types with different sizes covered by the annotators. The coverage of different topic types happened naturally, with no enforcement due to different interests of the annotators. While selecting their topics, annotators were allowed to see each others' profiles to prevent multiple annotation of the same topic. They were provided with example profiles and were encouraged to experiment with the system by creating and discarding trial profiles and annotations for learning purposes.

The annotation process begins with selecting a seminal event and finding its first story. For this purpose, annotators may choose an event that they remember, or identify an event by using Internet news portal archives or the information-retrieval facility of ETracker, which searches the news stories of the entire year of 2005 (Note that the annotations were done in early 2007 and that the test collection contains only the news of 2005.) For identifying the first story of a selected seminal event, annotators usually performed multiple searches over the corpus with queries using the ETracker's IR system. During this process, ETracker displays the "matching documents" in chronological order rather than in relevance order. Note that by careful query-term selection, such as named entities or event-related words, one can identify news articles related to a specific topic. The chronological display of

DOI: 10.1002/asi

documents helps annotators in finding the first story. After identifying the first story, annotators completed the topic profile. The correct selection of the first story is important. During the construction of BilCol-2005, the first story of each topic was approved by a senior annotator. If the first story of a topic was not approved, the process of selecting the first story and generating the associated event profile was repeated.

In ETracker, after identifying the first story, four annotation steps are performed for finding the tracking stories. These four steps are followed by a quality control performed by a senior annotator. In all steps, the listed documents are relavanced ranked with respect to the query used in that step, and they have a timestamp newer than that of the first story. To urge annotators to see the whole document, only document links are displayed without a snippet.

In all steps, the links of newly displayed documents are shown in blue. The links of the labeled stories are shown in red (if labeled as "off-topic"), green (if labeled as "on-topic"), or orange (if labeled as "off-topic" and "on-topic" at the same time). Annotators are allowed to read a story and change its label any number of times. The annotation steps are defined next.

1. *Search with the seed*: ETracker searches the collection for tracking stories by using the seed (i.e., first) story as a query. The annotator decides if the results are on topic, and labels them as "Yes: on-topic" or "No: off -topic." In this and the following steps, if the annotator is "unsure" about a story, he or she can mark it in both ways (i.e., "yes" and "no" at the same time and can change it later to "yes" or "no"). If a story remains marked like that after the completion of all steps, it is assumed as "off-topic." During the construction of BilCol-2005, there were a few such cases.
2. *Search with the profile information*: ETracker ranks by using the profile description words as a query.
3. *Search with on-topic stories*: ETracker takes the first three on-topic stories of Steps 1 and 2 and uses them as six separate queries (If they are not distinct, the following on-topic stories of each step are selected to gather six distinct stories, if possible). The final ranking of the stories retrieved by these queries is determined by using the reciprocal rank data-fusion method (Nuray & Can, 2006). They are displayed in the rank order determined by the data-fusion process.
4. *Search with queries*: ETracker employs the annotators' queries for finding a greater number of relevant stories. Annotators may use any number of queries.

The number of listed stories is limited to 200, 300, 400, and 200, respectively, for Steps 1 to 4. To make the annotation process more efficient and effective, there is a recommended time limit in Steps 1 to 4 as follows: 60, 45, 45, and 30 min, respectively. The number of labeled stories increases in the later annotation steps; therefore, in general, the time allotment per listed document decreased in later steps. Annotators can spend more time than the recommended time limits and can stop when they have reached the "off-topic threshold." Off-topic-threshold means that the last 10 stories evaluated by the annotator are off-topic, and the ratio of "the number of on-topic stories found so far" to "number of off-topic stories" is 1:2. This means that for example, if an annotator finds 10 on-topic stories in their list, they also must label at least 20 off-topic stories. At least the last 10 on the list must be labeled off-topic before they can move on to the next step of annotation. If annotators cannot find any on-topic stories in the top 50 stories of the first step, they are advised to drop the topic and try another one.

Quality control. A senior annotator examines at least 20 documents from each of the following categories: documents labeled as "on-topic;" documents labeled as "off-topic;" and documents listed, but not labeled. In this process, if any of the on-/off-topic is labeled incorrectly, or any document not examined is actually an on-topic document, then the junior annotator is asked to redo the annotations. In such cases, annotators are allowed to change the topic.

TABLE 2. News categories and number of annotated topics in each category.[a]

Category no.	News category	No. of topics
1	Elections	0
2	Scandals/Hearings	10
3	Legal/Criminal Cases	13
4	Natural Disasters	0
5	Accidents	16
6	Acts of Violence or War	11
7	Science and Discovery	4
8	Financial	2
9	New Laws	4
10	Sports	5
11	Political and Diplomatic Meetings	2
12	Celebrity/Human Interest	11
13	Miscellaneous	8

[a] Due to double category assignment to six topics, there are total of 86 topics.

BilCol-2005 Topic and Annotator Characteristics

The number of topics annotated in each category is shown in Table 2. Furthermore, some additional information about the collection and information about some sample topics are provided in Table 3. In Table 3, for the "Onur Air' in Avrupa'da yasaklanması (Banning of Onur Air in Europe)" topic, there are total of 159 tracking stories (i.e., "directly related" events and activities), and it stays active for 203 days; on the first 100 days of this topic, there are total of 154 related news stories. The topic annotation details can be seen in Öcalan (2009) and BilCol-2005 (2009).

The annotators who took place in the construction of BilCol-2005 are experienced Web users: graduate and undergraduate students, faculty members, and staff. They are not required to have an expertise on the topic that they pick. Altogether, there were 39 native-speaker annotators. They used interpretation rules that are similar to those used in the TDT

TABLE 3. Information about sample topics and some averages for BilCol-2005.

			No. of stories on first Fn days			
Sample topic (Topic No. in BilCol-2005)	No. of tracking stories	Life span (days)	Fn = 100	Fn = 50	Fn = 25	Fn = 10
Onur Air'in Avrupa'da yasaklanması (Banning of Onur Air in Europe) (2)	159	203	154	154	148	105
Londra metrosunda patlama (Explosion in London subway) (6)	454	175	440	419	376	236
400 koyun intihar etti (400 sheep commit suicide) (10)	10	8	10	10	10	10
Mortgage Türkiye'de (Mortgage has arrived in Turkey) (14)	375	357	60	41	25	13
Attilâ İlhan vefat etti (Attilâ İlhan passed away) (21)	40	70	40	37	36	32
Sahte rakı (Counterfeit rakı) (48)	323	182	316	291	255	197
İlk yüz nakli (First face transplantation) (61)	14	17	14	14	14	10
Averages for all 80 topics of BilCol-2005	73	92	64	54	47	36

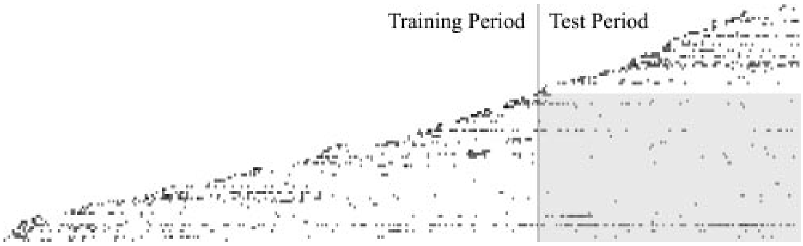


FIG. 2. The distribution of BilCol-2005 topic stories among the days of 2005. The *x* axis goes from January 1 to December 31, 2005; there are 80 topics, and each horizontal position on the *y* axis corresponds to a different topic. Each spot indicates the occurrence of one or more stories on a day. The gray level of spots is proportional to the number of stories on that day; darker spots indicate more stories. Days with 10 or more stories are shown with the same gray color. The lower right grayed segment shows the unused stories (Two lower topics with many stories in the gray area are used in testing). [The figure is blurry due to its nature and dimensions].

studies (TDT, 2004), except we allowed the annotators to select more than one category. However, a senior annotator inspected the topics to determine the quality of the annotations, including the interpretation of the coverage of the topic categories. Twenty-one low-quality events were deleted. In most of the eliminated topics, the related activities were not cohesive enough, and in some cases, the event coverage was incorrectly interpreted by annotators.

The final test collection contains 80 topics after the elimination of low-quality topics. On average, there are 73 (median = 32, minimum = 5, maximum = 454) tracking stories for each topic. On average, annotators spent, not counting breaks from work, 109 (median = 80, minimum = 20, maximum = 825) min for their annotations. The average topic life is 92 (median = 59, minimum = 1, maximum = 357) days. The distribution of topic stories among the days of 2005 is shown in Figure 2.

Evaluation Methodology

The most common evaluation measures in TDT are false alarm (FA) and miss rate (MR). Definitions of these effectiveness measures are as follows.

- FA = number of tracking stories labeled as new event/total number of tracking stories.
- MR = number of new events labeled as tracking stories/number of all new events.

These are both error measures, and the goal is to minimize them. In the ideal case, they are both equal to zero.

FA and MR are shown with a curve by using FAs and MRs gathered from various similarity threshold values that are used for decision making (Allan et al., 2000; Fiscus & Doddington, 2002). This curve is defined as a detection error trade-off (DET) curve, which is similar to the traditional receiver operating characteristic or relative operating characteristic curve (Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997). DET curves are plotted on a Gaussian (normal) deviate scale. The Gaussian deviant scale has advantages with respect to linear scales; for example, it expands the high-performance region (for more information, see Fiscus & Doddington, 2002, p. 24). DET curves provide a visualization of the trade-off between FA and MR. They are obtained by moving thresholds on the detection decision confidence scores. In obtaining the overall system performance, we use the topic-weighted approach that assigns the same importance to all topics, independent of their number of tracking stories. This approach is commonly used in the literature and are more preferable than story-weighted estimates (Fiscus & Doddington, 2002, p. 22). Numerical examples for story- and topic-weighted approaches can be seen in Baglıoğlu (2009, pp. 68–70).

DET curves provide detailed information; however, they may be difficult to use for comparison. For this reason, in TDT, there is another effectiveness measure—a detection cost function, C_{Det}—which combines FA and MR and yields a single value for measuring the effectiveness (Fiscus & Doddington, 2002). The detection cost function is defined as follows.

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot (1 - P_{target})$$

where

- $C_{Miss} = 1$ and $C_{FA} = 0.1$ are the costs of a missed detection and an FA, and they are prespecified;
- $P_{target} = 0.02$, the a priori probability of finding a target as specified by the application;
- P_{Miss}: miss probability (rate) determined by the evaluation result;
- P_{FA}: false alarm probability (rate) determined by the evaluation result.

The prespecified numerical values given previously are consistently used in TDT performance evaluation (Fiscus & Doddington, 2002; Fiscus & Wheatley, 2004, Manmatha, Feng, Allan, 2002).

The formula given for C_{Det} has a dynamic range of values and is difficult to use for relative comparison. For this reason, a normalized version of C_{Det} is used. In this case, C_{Det} is divided by the minimum expected cost obtained by either answering "yes" to all decisions or answering "no" to all decisions.

$$(C_{Det})_{Norm} = C_{Det}/\text{Minimum}[C_{Miss} \cdot P_{target}, C_{FA} \cdot (1 - P_{target})]$$

According to this formula, the value 0 reflects the best performance that can be achieved. The value 1 means a system is doing no better than consistently guessing "no" or "yes" (Fiscus & Doddington, 2002; Fiscus & Wheatley, 2004). In our work, we use the normalized version of the C_{Det} formula.

To evaluate performance, the stories are sorted according to their similarity scores, and a threshold sweep is performed with the similarity value increments of 0.001 beginning from 0.001 to the highest possible value. The parameter sweep approach is used in all training experiments reported in this study. For example, during NED, all stories with scores below the NED threshold θ_{on} are declared as new; other cases are treated as old. At each threshold value, the MRs and FAs are identified, and a cost is calculated as a linear function of their number. The threshold that results in the least cost is selected as the NED threshold (Kumaran, Allan, & McCallum, 2004). Different NED systems are compared based on their minimum cost. This minimum cost is defined as Min. $C_{Det} = \min\{(C_{Det})_{Norm}\}$ where $(C_{Det})_{Norm} \in S$, and S is the set of all minimum normalized cost values calculated by performing a threshold sweep. A similar approach is used in TT.

Moreover, we also used statistical one-tailed paired t tests over topics' C_{Det} value using $\alpha = 0.05$ for the signficance level. In the statistical test, the difference between pair samples are assumed to be a random sample from a normal distribution, with mean zero and unknown variance against the alternative that the mean is not zero. Before doing the statistical tests, the C_{Det} values, which were identified as potential outliers, were eliminated. A potential outlier is defined as an observation that is more than 2.5 *SD*s above or below the mean. In other words, we first compute the differences for each pair and then extract the pair whose value is more or less than 2.5 *SD*s from the mean. This approach is repeated with the remaining observations. We observed only one and two outliers, respectively, in two cases in our first attempt.

NED and TT Methods Used in the Study

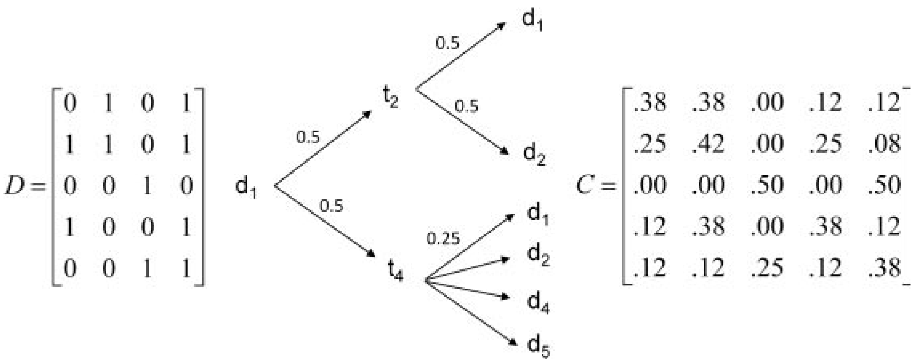
Our NED and TT methods are similar to the ones available in the literature (e.g., for NED, see Allan, Lavrenko, & Connell, 2003; for adaptive TT, see Allan, Papka, & Lavrenko, 1998; Leek, Schwartz, & Sista, 2002). We use the sliding time-window concept for NED. During NED, we compare the newest story with the time-window stories; if the newcomer is different (i.e., dissimilar) enough from them, this condition is defined as NED, and it is treated as the first story of a new event. During TT, for a given topic, the newcomer is compared with the topic description vector; if it is similar enough, it is assumed that it is a tracking story for that topic. During TT, each topic is handled separately.

Document Indexing: Incremental idf Approach

For the calculation of the similarity values, each story is represented by a document vector of size n using its dn number of terms with the highest *tf.idf* scores. Here, n indicates the number of unique terms that appear in the collection so far ($dn \ll n$). By using the terms with the highest *tf.idf* values (Salton & Buckley, 1988), we aim to index documents by using their most important terms. We refer to dn as "document vector length." Note that actual document vector length of some documents can be smaller than dn since they may not contain that many number of unique words. The *tf.idf* formula is defined as follows.

$$w(t, \vec{d}) = (1 + \log_2 tf(t, \vec{d})) \cdot \log_2(N_t/n_t)$$

In this formula, $w(t, \vec{d})$ is the weight of term t in document (vector) $\vec{d}$; $tf(t, \vec{d})$ is the number of occurrences of term t in document d; $\log_2(N_t/n_t)$ is the *idf* (inverse document frequency) component of the formula, where n_t is the number of stories in the collection that contains one or more occurrence of term t including the newcomer, and N_t is the number of accumulated stories so far in the collection. Hence, n_t and N_t (and therefore the *idf* values) are incrementally computed. A similar approach has been used in other studies (e.g., Yang et al., 1998). We use an auxiliary corpus containing the 2001 to 2004 news stories, about 325,000 documents from *Milliyet Gazetesi* (Can et al., 2008b). We use this retrospective corpus to obtain *idf* statistics, and update the *idf* values with each incoming story. This term-weigthing method is used with the cosine similarity measure. In a later section, the same term-weighting formula also is used with the Dice, Jaccard, and



$$D=\begin{bmatrix}0&1&0&1\\1&1&0&1\\0&0&1&0\\1&0&0&1\\0&0&1&1\end{bmatrix}\qquad C=\begin{bmatrix}.38&.38&.00&.12&.12\\.25&.42&.00&.25&.08\\.00&.00&.50&.00&.50\\.12&.38&.00&.38&.12\\.12&.12&.25&.12&.38\end{bmatrix}$$

FIG. 3. From left to right: example binary D matrix ($m=5$, $n=4$), hierarchical representation of the two-stage probability model for d_i of the D matrix, and C: cover coefficient matrix (Some values are approximate).

overlap similarity measures. The cover coefficient (CC) similarity measure has its own term-weighting approach. Later, we use the Hellinger and Okapi similarity measures, and they also have their own term-weighting formulas.

Similarity Measures

Cosine similarity measure. In the vector space model (Salton, 1989), the cosine similarity measure is the cosine of the angle between two vectors in an n-dimensional space. According to the cosine measure, the similarity between two documents d_i and d_j is defined as follows (Salton, 1989; Salton & Buckley, 1988; van Rijsbergen, 1979).

$$sim(d_i,d_j)=\left[\sum_{k=1}^{n} w_{ik}\cdot w_{jk}\right]\cdot\left[\sum_{k=1}^{n} w_{ik}^2\cdot\sum_{k=1}^{n} w_{jk}^2\right]^{-1/2}$$

Here, n indicates the document vector size; w_{ik} and w_{jk} indicate the *tf.idf* weights of term-k (t_k) in documents d_i and d_j, respectively.

CC-based similarity measure. According to the CC concept, the coverage (defined later) of d_i by d_j, based on a document by term D matrix of dimensions m and n, is expressed as follows (Can, 1993; Can & Ozkarahan, 1990).

$$c_{ij}=\sum_{k=1}^{n}[\alpha_i\cdot d_{ik}]\cdot\left[\beta_k\cdot d_{jk}\right]$$

where

$$\alpha_i=\left[\sum_{l=1}^{n} d_{il}\right]^{-1},\ \beta_k=\left[\sum_{l=1}^{m} d_{lk}\right]^{-1},\ 1\le i\le m, 1\le j\le m$$

In this formula, d_{ik} indicates the number of occurrences of term t_k in document d_i (A similar definition applies to d_{jk}, etc.); α_i and β_j are the inverse of row i and column j sums of D. The symbol c_{ij} indicates the coverage of d_i by d_j (i.e., the probability of selecting any term of d_i from d_j). It can be interpreted as an asymmetric similarity measure: If document vectors of d_i and d_j are different, then c_{ij} and c_{ji} have different values.

Obtaining the value of c_{ij} involves a two-stage probability experiment (Hodges & Lehmann, p. 94): In the first stage, we randomly choose a term t_k of document d_i (indicated by the first product term of the c_{ij} formula); in the second stage, we randomly choose t_k from document d_j (indicated by the second product term of the c_{ij} formula). Figure 3 illustrates the concept with a binary D matrix (For simplicity, a binary matrix is preferred.) The element c_{ij} for $i\neq j$ indicates the extent to which d_i is covered by d_j (or coupling of d_i with d_j), and for $i=j$, it indicates the extent to which d_i is covered by itself (decoupling of d_i from the rest of the documents). The row sums of the C matrix are equal to 1; c_{ii} is equal to 1 if d_i is completely decoupled from the rest of the collection (i.e., if its terms do not appear in any other document) (for a detailed explanation of the CC concept, see Can & Ozkarahan, 1990; Yu & Meng, 1998).

Within the context of our problem area, we assume that D is an abstraction for the members of the sliding time window. In terms of Figure 3, we assume that d_1 is the most recent document and that d_2 to d_5 are the old members of the time window. In the example D matrix, document vector size (n) is 4, "document vector length" (dn) of d_1 is 2. The tree structure of Figure 3 shows the hierarchical representation of the two-stage probability model for d_1, and C matrix gives the CC values among all documents. Some similarities (e.g., both are asymmetric) can be drawn between the CC and KL divergence measures (Manning, Raghavan, & Schütze, 2008); however, such concerns are beyond the scope of this study.

In the experiments, we tried various options for finding the relationship between the newest document and the individual members of the sliding time window by using the CC concept; for example, by (a) only considering the time-window documents in the calculation of the β values, (b) using an incremental approach with a retrospective document collection (2001–2004 news stories of *Milliyet Gazetesi*) by assuming that the arriving documents are added into this retrospective collection in the calculation of the β values,

1. Obtain the terms of d.
2. Update the idf statistics using the terms of d.
3. Select the highest $tf.idf$ weighted dn number of terms of d for its vector representation.
4. Compute the similarity between d and the time-window stories; obtain the confidence score of d, $\max_{d_k \in window}(sim(d, d_k))$ i.e., the maximum similarity value among the computed values.
5. If the confidence score is less than the NED threshold θ_{on} (θ_{CC} for CC and θ_{cosine} for cosine), then d is labeled as the first story of a new event.

FIG. 4. New event detection (NED) cosine and cover coefficient (CC) methods.

and (c) using a variant of the incremental approach by modifying the values α_i and β_k as follows for smoothing their effects on c_{ij}. (We assume that each summation gives a value greater than 1.)

$$\alpha_i = \left[\log \sum_{l=1}^{n} d_{il}\right]^{-1}, \beta_k = \left[\log \sum_{l=1}^{m} d_{lk}\right]^{-1}$$

In this approach, we take logarithms to not to overly diminish the d_{ik} and d_{jk} (*tf*) values in the c_{ij} formula since the original α_i and especially β_k values can be too large. Compared to the original c_{ij} formula, the logarithmic approach assigns more emphasis to the *tf* values of the members of the sliding time window. Similar normalization approaches are used in other similarity measures such as Inquery (Kowalski, 1997, pp. 104, 116). In the experiments, the best results with CC are obtained with the logarithmic approach; in this article, we present the results associated with this approach.

Vural (2002) also used a CC-based concept called "cluster seed power" in TDT. In his study, if a newcomer has the "power" of becoming a cluster seed (Anderberg, 1973, pp. 157–159), then it is selected as the first story of a new event. The details are beyond the scope of this work (see Vural, 2002).

NED Methods

In the sliding time window (see Figure 1), we only keep the stories of a certain number of most recent days and compare the newcomer to them (Luo et al., 2007; Papka, 1999; Yang et al., 1998).

Stand-alone use of similarity measures for NED. In this approach, the flag of "first story" is assigned to the newest story d if its confidence score is below a predetermined threshold, where the confidence score of d is defined as $\max_{d_k \in window}(sim(d, d_k))$ (i.e., the maximum similarity observed between d and the members of the time window). The NED method is defined in Figure 4.

The predetermined thresholds for the cosine and the CC measures, namely θ_{cosine} and θ_{CC}, are obtained by training; these values provide the Min. C_{Det} values for the respective measures.

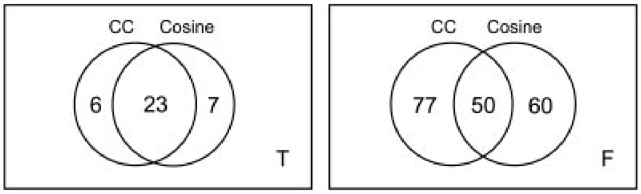


FIG. 5. New event detection (NED) consistency of cover coefficient (CC) and cosine decisions during training for 50 events: Left Venn diagram is for true matches; right Venn diagram is for false alarms.

Combination methods: Experimental evidence for our intuition. In this article, we hypothesize that decisions of two different similarity measures can be combined to lower the cost. First, we provide experimental evidence that supports this expectation (Note that the results reported here are obtained by the most favorable conditions as defined in the section "New Event Detection: Experimental Results" using the training dataset defined in the "Experimental Environment" section.) For this purpose, we analyzed the NED decision consistency of the CC and cosine similarity measures. The left Venn diagram of Figure 5 shows consistency in the correct NED decisions (true matches or hits) when the threshold values obtained during training are used on the same dataset for decision making. It shows that these two measures agree on 23 decisions jointly, 7 cases are only detected by cosine, 6 cases are only detected by CC; altogether, we have 36 correct decisions of 50. The consistency between CC and cosine decisions in correct decisions is about 78% (23/30 and 23/29). The right Venn diagram shows the case for the FAs generated by tracking news stories during NED. It shows that the cosine and CC measures incorrectly detect 110 and 127 tracking stories, respectively, as new events, and among these, they have 50 common false drops. Their agreement in making wrong decisions together is about 42%.

The observations indicate that these measures are more consistent in correct decisions and less consistent in wrong decisions. This intuitively implies that they potentially may correct each other's incorrect decisions if they are used together for decision making: (a) The left illustration suggests that when cosine "or" CC detects a new event, and if we accept that decision, we may be able to "increase" the number of correct new-event detections (i.e., decrease miss rate); (b) the right illustration suggests that if cosine "and" CC decisions agree, and if we make a decision when both of them agree, then we may be able to "decrease" the number of incorrect new-event detections (i.e., decrease FA rate).

And- and or-combination of similarity measures for NED. In these methods, we combine the cosine and CC similarity scores in two different ways by changing Step 5 of Figure 4

1. Obtain the terms of d.
2. Update the idf statistics using the terms of d.
3. Select the highest $tf.idf$ weighted dn number of terms of d for is vector representation.
4. Compute the confidence score: $sim(T, d)$, the similarity between the current story (d) and the target topic T.
5. If the confidence score is above the TT threshold θ_{on} (θ_{CC} for CC and θ_{cosine} for cosine, obtained by training), then d is classified as an on-topic story, otherwise it is classified as off-topic.
6. Only for adaptive TT: If d is classified as on-topic and if θ_{on} is above the pre-determined adaptation threshold θ_{adapt} (different θ_{adapt} values are used for cosine and CC) then update the vector representation (centroid) of T using existing centroid terms and the terms of document d.

FIG. 6. Static and adaptive topic tracking (TT)—cosine and cover coefficient (CC) methods.

as follows by using one of the combination alternatives given next.

- And-combination: If $x < \theta_{cosine}$ *and* $y < \theta_{CC}$, then d is labeled as the first story of a new event.
- Or-combination: If $x < \theta_{cosine}$ *or* $y < \theta_{CC}$, then d is labeled as the first story of a new event.
  where x and y, respectively, indicate the confidence score of CC and cosine; that is, $x = \max_{d_k \in window}(sim_{cosine}(d, d_k))$ and $y = \max_{d_k \in window}(sim_{cc}(d, d_k))$.

TT Methods

The TT task uses a few (usually one to four) sample stories about a given topic T and aims to find stories on that topic ("on-topic" stories) in a news stream. In TT, there would be several topics tracked at the same time, and each one is tracked separately and individually. This means that the decision made for one topic does not affect decisions made for the other topics (TDT, 2002); and unlike information filtering, it typically involves no user feedback (Allan, 2002b).

Stand-alone use of similarity measures for TT. In TT, the definition of a topic T is done by using a sample document vector, and each incoming news-stream story (d) is considered one by one. The definitions of the static and adaptive TT methods are provided in Figure 6. Note that unlike the study of Elsayed and Oard (2005), in our work the term *adaptive* implies no user interaction or human supervision. In adaptive TT, for on-topic documents if the similarity score is above the threshold θ_{adapt} (obtained by training), the vector representation of T is updated using its current terms and those of document d. By using this additional similarity score (θ_{adapt}), not all tracking documents but only highly similar tracking documents are used to modify the topic description vector. A similar approach can be seen in Leek et al. (2002). In adaptive tracking, topic-description vectors are treated as topic centroids, similar to the ones used in cluster-based retrieval (Altingovde, Demir, Can, & Ulusoy, 2008; Can, Altingovde, & Demir, 2004; van Rijsbergen, 1979). During this process, the highest weighted dn number of terms in the story are added to the topic centroid, and then the topic centroid is redefined by selecting its highest weighted dn number of terms. This approach aims to remember old words and also follow topic changes by focusing on new developments.

TABLE 4. Distributions of stories among training and test sets.

Corpus	Time span (month.day.year)	No. of topics	No. of documents	No. of tracked documents
Training	01.01.2005– 08.31.2005	50	141,910	3,358
Test	09.01.2005– 12.31.2005	32	67,395	2,288

And- and or-combination of similarity measures for TT. In these methods, we combine the cosine and CC similarity scores in two different ways by changing Step 5 of Figure 6 as follows (In the following, x and y indicate the CC and cosine similarity, respectively, of the incoming story to the topic under consideration.)

- And-combination: If $x > \theta_{cosine}$ *and* $y > \theta_{CC}$, then the incoming story d is labeled as a tracking story of T.
- Or-combination: If $x > \theta_{cosine}$ *or* $y > \theta_{CC}$, then the incoming story d is labeled as a tracking story of T.

Experimental Environment

Dividing BilCol-2005 Into Training and Test Sets

For experimental evaluation, we divide the BilCol-2005 test collection into training and test sets. For this purpose, the news stories of the first 8 and last 4 months serve as the training and test collections, respectively. The division gives us the opportunity of keeping most of the tracking stories together with their corresponding first stories. For example, dividing the dataset into two 6-month periods (January–June and July–December) does not give us that opportunity (see Figure 2). Altogether, we have 80 topics. For two topics (Bilcol-2005, 2009; Topics 14 and 15) used for training, there is a considerable number of news stories in the period that corresponds to the test data (In Figure 2, they are two lower stories beginning on January 7 and 15.) For these two topics, their first stories in the test-set section are used as the first stories of the two new events. Using this approach, there are 82 topics altogether in the training and test sets (i.e., two more than the original 80 topics). Although using nonoverlapping stories of two topics both in training and test sets is a minor point, this can be an issue in some approaches such as the ones that involve learning topic models (Yi & Allan, 2008). As can be seen from Table 4, the average number of news stories per topic in the training and test sets are approximately the same: 67.16 (3,358/50) and 71.50 (2,288/32), respectively, stories.

TABLE 5. Min. C_{Det} values for cover coefficient (CC) and cosine measures for new event detection (NED) with various document vector length (*dn*) and stemmer combinations[a] (using the longest stopword list).

dn (doc. vector length)	CC effectiveness				Cosine effectiveness			
	NS	F5	F6	LM	NS	F5	F6	LM
10	0.8996	0.8957	0.8863	0.8743	0.9517	0.7683	0.7555	0.7590
20	0.8092	0.7292	0.7462	0.6842	0.7669	0.7000	0.7437	0.6204
30	0.7268	0.6657	0.6408	0.6751	0.7476	0.6352	0.7115	0.6220
40	0.6849	0.6714	0.6198	0.6433	0.7222	0.6284	0.6562	0.5977
50	0.6841	0.6361	0.6097	**0.5950**	0.7696	0.6455	0.6485	0.5860
60	0.6411	**0.5929**	**<u>0.5730</u>**[a]	0.5982	0.7695	0.6312	0.6771	0.5974
70	0.6325	0.6358	0.5926	0.6235	0.7635	0.6277	0.6692	0.6019
80	0.6300	0.6490	0.5973	0.6740	0.7274	0.6354	0.6581	0.6115
90	0.6255	0.6462	0.6151	0.7408	0.7248	0.6349	0.6405	0.5916
100	**0.6043**	0.6512	0.6586	0.7610	0.7025	0.6347	0.6337	0.5849
120	0.6325	0.7306	0.6792	0.8001	0.7007	0.6326	0.6661	0.5972
140	0.6686	0.7637	0.7482	0.8142	0.7210	0.6357	0.6572	0.5913
160	0.7093	0.7639	0.7599	0.8542	0.7411	0.6313	0.6516	0.5965
180	0.7570	0.7872	0.7610	0.8658	0.7225	0.6313	0.6504	0.5912
200	0.7917	0.7857	0.7728	0.8746	0.7155	0.6331	0.6458	0.5909
All terms	0.8012	0.8943	0.8512	0.9444	**0.6872**	**0.6101**	**0.6176**	**<u>0.5777</u>**[a]

[a] Best case for each stemmer is in bold, and best case of each similarity measure is underlined.

Stemming Methods Used in the Study

In our recent work, it was shown that stemming has a significant effect on Turkish IR (Can et al., 2008b). Therefore, in this study, we examine the effects of stemming on Turkish TDT. In this study, we use three stemming methods in obtaining vectors used for document description: (a) no stemming, so called "austrich algorithm;" (b) first-n, n-prefix, characters of each word—two versions, n = 5 and n = 6- and (c) a lemmatizer-based stemmer.

- *No-Stemming* (NS): The NS option uses words as they are as an indexing term.
- *Fixed Prefix Stemming* (F5, F6): The fixed prefix approach is a pseudo stemming technique. In this method, we simply truncate the words and use the first-n (Fn) characters of each word as its stem; shorter words are used as is. This approach also can be interpreted as a restricted case of character n-grams. In this study, we experiment with F5 and F6, which have been experimentally shown to give the best performance in Turkish IR (Can et al., 2008b).
- *Lemmatizer-Based Stemming* (LM): A lemmatizer is a morphological analyzer that examines inflected word forms and returns their dictionary forms (e.g., "good" is returned for "best"). Lemmatizers are not stemmers since the latter tries to find a common base form for a word using a heuristic process by dropping the ends of words; in contrast, a lemmatizer aims to find the dictionary entry of a word with the use of a vocabulary and morphological analysis of words. In this article, we prefer the word "stemming" over lemmatization as it is more commonly used.

The fixed prefix approach is a simple method that provides an IR performance similar to that of a complicated lemmatizer-based stemmer (Can et al., 2008b). In this work, we want to see if the same also is true in TDT applications. NS provides a baseline for comparison.

New Event Detection: Experimental Results

Document Vector Length, Stemmer, and Window-Size Considerations

During NED, numerous combinations are possible for the document vector length (*dn*), stemmer (F5, F6, LM, NS), and window-size parameters. In general, it would be reasonable to choose a window size that would give an opportunity of finding the tracking stories of topics. For most cases, the average time distance among the stories of a topic is less than or equal to 12 days. To select a suitable "document vector length and stemmer" combination, we used the 12-day sliding time-window size and analyzed the system effectiveness (performance measured in terms of Min. C_{Det} values). Table 5 shows the effectiveness in terms of Min. C_{Det} values with the CC and cosine similarity measures with different document vector length (*dn*) and stemmer (NS, F5, F6, LM) combinations (In the experiments of this section, we used the longest stoplist. More detail on stoplist selection is provided in the next section.)

In general, F5, F6, and LM stemming approaches provided better performance than did the no stemming (NS) approaches. For example, with the cosine measure, these stemmers outperformed NS in all document vector length cases.

With the cosine measure, the combination "all terms and LM" provided the best performance; furthermore, in 15 of the 16 different document vector lengths, LM provided the best performance with respect to the other stemmers. For all stemmer cases, using all terms gave the lowest Min. C_{Det} value (i.e., the best performance). With the cosine measure, observing the best performance with all terms is consistent with the results of other researchers (e.g., Allan, Lavrenko, & Swan, 2002, p. 200).

TABLE 6. Effects of stemmer on new event detection (NED) effectiveness with cover coefficient (CC) (with $dn = 60$ terms) and cosine (with all terms) measures. Best cases are in bold.

	CC			Cosine		
	Training results			Training results		
Stemmer	Min. C_{Det}	Threshold	Test results C_{Det}	Min. C_{Det}	Threshold	Test results C_{Det}
NS	0.6411	0.199	0.7731	0.6872	0.099	0.7280
F5	0.5929	0.252	0.7637	0.6101	0.166	0.6060
F6	**0.5730**	0.254	**0.6476**	0.6176	0.155	0.6627
LM	0.5982	0.290	0.7672	**0.5777**	0.200	**0.4947**

TABLE 7. Effects of stopword list size on new event detection (NED) effectiveness with cover coefficient (CC) and cosine similarity measures.

	CC			Cosine		
	Training results			Training results		
Stopword list size	Min. C_{Det}	Threshold	Test results C_{Det}	Min. C_{Det}	Threshold	Test results C_{Det}
0	0.7149	0.256	0.7622	0.7050	0.226	0.5809
10	0.6512	0.246	0.7199	0.7052	0.211	0.5870
147	0.5895	0.244	0.7286	0.5958	0.200	0.5292
217	0.5730	0.254	0.6476	0.5777	0.200	0.4947

In terms of the CC measure, F6 with the document vector length of 60 provided the best performance. In fact, F6 yielded the best performance (i.e., the lowest Min. C_{Det} value) for most document vector lengths: In 10 cases of 16 different vector lengths, F6 provided a lower cost than that of NS; in 14 cases, it is better than F5; and in 13 cases, it is better than LM.

Now, we focus on the conditions of these best performances. We hypothesize that with the CC similarity measure, the use of F6 with a document vector length of 60 would provide a statistically significantly higher effectiveness than that of using other stemming approaches. Likewise, we hypothesize that with the cosine similarity measure, the use of LM when all document terms are used in obtaining document description vectors would provide a statistically higher effectiveness than would other stemming approaches. The training and test results for these cases with all stemmers are provided in Table 6 (Test results were obtained by using the similarity threshold values which are obtained by training.) The one-tailed statistical test results (using the test set) show that when cosine is used, the C_{Det} value of LM is significantly smaller than those of NS ($p \leq 0.001$) and F5 and F6 ($p \leq 0.05$). When CC is used, the C_{Det} value of F6 is significantly smaller than those of NS and F5 ($p \leq 0.05$), but it is slightly smaller than that of LM ($p \leq 0.10$). The F6 versus LM results with CC show that in Turkish NED, depending on the used similarity measure, a simple word truncation stemming method can compete with a lemmatizer-based stemmer.

In the rest of the article, unless otherwise specified, we use the stemmer and document vector length combinations of "F6-60 terms" for CC, and "LM-all terms" for cosine.

Word Stopping Strategy Considerations

In IR, a stoplist contains frequent words that are ineffective in distinguishing documents from each other. Elimination of such words can improve effectiveness in IR (Croft, Metzler, & Strohman, 2009). Kumaran and Allan (2004) showed positive influence of category-based word stopping on NED. We investigate the effects of word stopping by using four stoplists with different sizes. As a baseline, we use a stoplist that contains no words. This is followed by a stoplist that contains 10 frequent words, as determined in our recent work on Turkish IR (Can et al., 2008b). We also experiment with two additional semiautomatically generated stoplists containing 147 and 217 words. The larger sets are inclusive of the smaller ones. The stoplist generation approach is defined in Can et al. (2008b). The first two lists are provided in Can et al. (2008b), and all of them are available in Kardaş (2009).

The results presented in Table 7 show that stoplists and their sizes have an influence on effectiveness. As we increase the stoplist size, positive influence of word stopping with respect to the baseline, no stopping, tend to increase (Note that in these experiments, time-window size, stemmer, and document vector length are kept the same; only the stoplist varies. The training Min. C_{Det} values of the last row of Table 7 also can be seen in Table 5. The experimental conditions that are specified for Table 5 also are used to obtain the C_{Det} values of Table 7.)

We also conducted one-tailed paired t tests on the Min. C_{Det} values of the training set and C_{Det} values of the test set. The statistical results on the training and test corpus indicate that using the longest stoplist yields Min. C_{Det} and C_{Det} values

TABLE 8. New event detection (NED) training results (Best stand-alone and combination results are in bold, and best case is underlined).

Similarity measure–combination method	Miss rate	False alarm	Min. C_{Det}	NED threshold (θ_{on})
CC	0.4200	0.0285	**0.5599**	$\theta_{CC} = 0.254$
cosine	0.4200	0.0322	0.5777	$\theta_{cosine} = 0.200$
and-combination	0.5600	0.0136	0.6268	$\theta_{CC} = 0.254$ $\theta_{cosine} = 0.200$
or-combination	0.2800	0.0471	**0.5108**	$\theta_{CC} = 0.254$ $\theta_{cosine} = 0.200$

that are significantly smaller than using no stopwords for both similarity measures ($p \leq 0.05$). As a result, the use of such stopwords increases NED effectiveness. On the other hand, there is no statistically significant difference between using the longest stoplist and using the list with 147 words, but we use the longest list in the remaining experiments because it yields the lowest cost.

The positive influence of word stopping can be explained by the fact that such words frequently appear in news stories, and during NED, news articles are used like a query. In IR, it is argued that improvements because of stopword use are due to the improper relative weight of *idf* in the term-weighting formula (Dolamic & Savoy, in press). Stopword-related issues can be further investigated within the context of TDT (e.g., see Kumaran & Allan, 2004).

Reassessing Window Size

After these observations, we reassessed the effect of the time-window size and obtained the Min. C_{Det} values with several different time-window sizes with the training dataset (Kardaş, 2009, Figure 4.2). The results show that the window size of 12 days is the best choice for the cosine similarity measure since this window size provides the lowest Min. C_{Det} values with this similarity measure (Actually for cosine, 12 and 18 days provide the same performance; for efficiency, we prefer a smaller—12 days—window size.) For the CC similarity measure, the window size with the lowest Min. C_{Det} value is determined as 14 days. In the rest of the article, we use these window sizes for the cosine (12 days) and CC (14 days) similarity measures. However, one can argue that the window size to be used should be independent of the similarity measures since it is a characteristic of the data. The small difference can be attributed to noise in data.

Comparative Evaluation of NED Methods

Training results. Figure 7 gives the new event detection DET plots (NIST, 2008) for CC and cosine. The figure shows that at smaller FA values, CC provides lower MRs; for FA values greater than about 7 (%), cosine provides lower MRs.

The optimum training results for each new event detection approach are given in Table 8. The optimum similarity thresholds values for the CC and cosine measures are 0.254 and 0.200, respectively. In addition, we observe that optimum thresholds for the and- and or-combination methods are the same as the case with the stand-alone use of these similarity measures. The best performance with the stand-alone and combined versions are in bold. The MRs are much higher than the FAs. When the similarity measures are used alone, CC provides a slightly smaller cost value (i.e., better effectiveness) than that of cosine (0.5599 vs. 0.5777); however, it is not statistically significant. It can be observed that in training, the or-combination gives the best performance in NED since it misses a lesser number of new events. This can be explained by the fact that a new event missed by CC can be caught by cosine, or vice versa. The or-combination's cost (0.5108) is significantly smaller ($p = 0.0493$) than that of the and-combination's (0.6268).

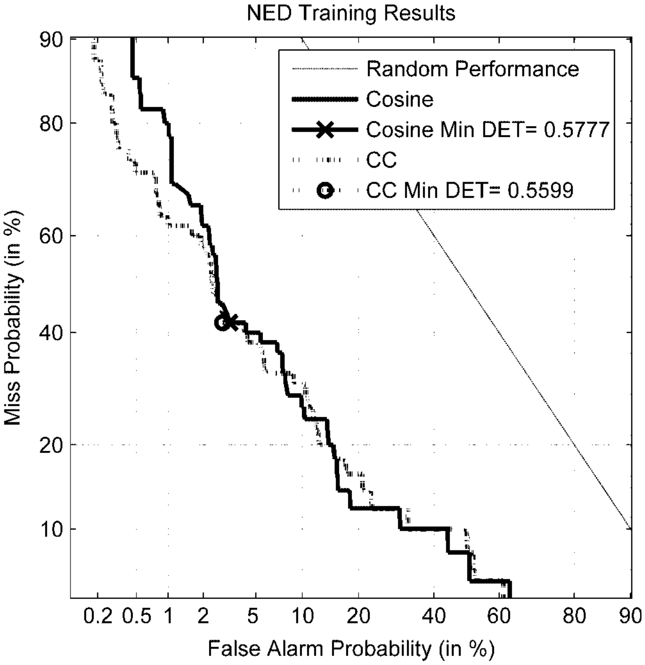


FIG. 7. New event detection (NED) training performance with cover coefficient (CC) and cosine similarity measures.

Test results. The test results are presented in Table 9. As in training, the MRs are much higher than the FAs. As expected, the or-combination method is effective in lowering MRs. During testing, the or-combination method provides the best performance, and results in percentage reduced costs of −8.03% (0.4550 vs. 0.4947) and −28.45% (0.4550 vs. 0.6359) with respect to the stand-alone use of cosine and CC,

TABLE 9. New event detection (NED) test results (Best stand-alone and combination results are in bold, and best case is underlined).

Method	Miss rate	False alarm	C_{Det}
Cover coefficient	0.5313	0.0214	0.6359
Cosine	0.3750	0.0244	**0.4947**
and-combination	0.6250	0.0103	0.6757
or-combination	0.2813	0.0355	**0.4550**

respectively. Or-combination C_{Det} values are significantly smaller than those of the and-combination ($p = 0.0277$). However, there is no statistical evidence supporting the difference between the or-combination and cosine; the same also is true for the or-combination and CC. During training, the cost of CC (0.5599) is smaller than that of the cosine (0.5777), but it is not statistically significantly smaller. The test results contradict training results: During testing, the cost of cosine (0.4947) is smaller than that of CC (0.6359). This is an impressive 22% difference, but it is not statistically significantly smaller. The contradiction of traning and test results can be explained by the fact that both training and test results are not statistically significant. In general, the or-combination method is more commendable than are the others: It decreases the MR (or-combination's MR is 33% lower than those of CC and cosine, 50% lower than that of the and-combination). As a result of the decrease in MRs, the or-combination also yields the lowest cost in terms of C_{Det} value.

Topic Tracking: Experimental Results

To start TT, some sample stories are needed about the topic to be tracked. In the literature, usually between one and four documents are used as sample stories. In our case, we use only the first story to obtain the topic description vectors. In static TT, these vectors remain unchanged. In adaptive TT, as explained earlier in the "methods" section, topic description vectors are treated as *topic centroids* and are updated according to the newly tracked documents.

The experiments show that for the CC measure, using the highest *tf.idf* weighted 60 terms for the description of news stories and topics provides the best performance (60 = 50% of the average number of unique nonstopwords per document). This is true both for static and adaptive TT. For the cosine measure, in static TT, the use of all terms of sample topic documents gives the best performance. In adaptive tracking, using 100 terms with the highest *tf.idf* weights gives the best performance (100 = 83% of the average number of unique nonstopwords per document). Connell et al. (2004) used a similar approach for adaptive tracking.

Comparative Evaluation of TT Methods

Training results. Figure 8 gives the static TT DET plots for CC and cosine. A comparison of Figure 8 with Figure 7 shows that TT is considerably more effective than is NED.

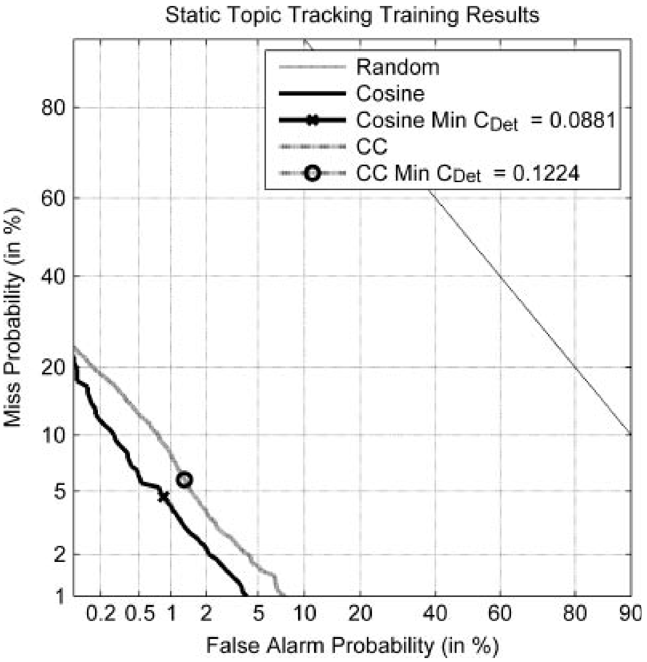


FIG. 8. Static topic tracking (TT) training performance with cover coefficient (CC) and cosine similarity measures.

In Figure 8, the difference between the CC and cosine performances is noticeable. Min. C_{Det} value for cosine (0.0881) is statistically significantly smaller ($p = 0.0092$) than that of CC (0.1224).

The training results (in terms of Min. C_{Det} and threshold values) of all methods both for static and adaptive TT are presented in Table 10. In adaptive tracking, the adaptation threshold (θ_{adapt}) is determined after obtaining the thresholds for CC and cosine. In static TT, the and-combination provides the lowest cost. Furthermore, its Min. C_{Det} value is statistically significantly smaller than those of the other static TT approaches (for all cases, $p \leq 0.05$). The or-combination Min. C_{Det} value is significantly smaller than that of CC ($p = 0.092$); however, there is no significant difference between the or-combination and cosine.

The adaptive topic tracking Min. C_{Det} values are better (i.e., lower) than those of static TT. Similar to the static case, the and-combination provides the best effectiveness. Its Min. C_{Det} value is statistically significantly smaller than those of CC and cosine (for both cases, $p \leq 0.05$). It is close to being statistically significantly smaller than that of the or-combination. For this case, we have suggestive ($p = 0.0902$) evidence in that direction.

Test results. The test results for static and adaptive TT are provided in Table 11. The table shows that in static TT, the and-combination and CC, respectively, provide the highest (lowest cost: 0.0824) and lowest effectiveness (highest cost: 0.1277) scores. The statistical test results show that the and-combination significantly improves the effectiveness with respect to CC, cosine, and or-combination methods (for all cases, $p < 0.05$). As in the training case, cosine provides

TABLE 10. Static and adaptive topic tracking (TT) training results (Best stand-alone and combination results are in bold, and best case of static and adaptive results is underlined).

TT submethod	TT method	Miss rate	False alarm	Min. C_{Det}	TT threshold (θ_{on})	Adaptive threshold (θ_{adapt})
Static	CC	0.0580	0.0130	0.1224	$\theta_{CC} = 0.097$	N/A
	Cosine	0.0462	0.0085	**0.0881**	$\theta_{cosine} = 0.099$	N/A
	And-combination	0.0425	0.0083	**0.0824**	$\theta_{CC} = 0.045$ $\theta_{cosine} = 0.094$	N/A
	Or-combination	0.0462	0.0085	0.0881	$\theta_{CC} = 0.335$ $\theta_{cosine} = 0.099$	N/A
Adaptive	CC	0.0540	0.008	0.0910	$\theta_{CC} = 0.085$	$\theta_{CC} = 0.400$
	Cosine	0.0320	0.007	**0.0650**	$\theta_{cosine} = 0.076$	$\theta_{cosine} = 0.300$
	And-combination	0.0260	0.007	**0.0580**	$\theta_{CC} = 0.027$ $\theta_{cosine} = 0.069$	$\theta_{CC} = 0.400$ $\theta_{cosine} = 0.300$
	Or-combination	0.0330	0.006	0.0650	$\theta_{CC} = 0.149$ $\theta_{cosine} = 0.080$	$\theta_{CC} = 0.400$ $\theta_{cosine} = 0.300$

CC = cover coefficient; N/A = not applicable.

TABLE 11. Static and adaptive topic tracking (TT) test results (Best stand-alone and combination results are in bold, and best static and adaptive case is underlined).

	TT submethod					
	Static TT			Adaptive TT		
TT method	Miss rate	False alarm	C_{Det}	Miss rate	False alarm	C_{Det}
CC	0.0355	0.0188	0.1277	0.0580	0.003	0.0699
Cosine	0.0494	0.0071	**0.0842**	0.0350	0.004	**0.0563**
And-combination	0.0445	0.0071	**0.0791**	0.0200	0.006	0.0513
Or-combination	0.0494	0.0071	0.0843	0.0170	0.006	**0.0461**

CC = cover coefficient.

an effectiveness (0.0842) statistically significantly smaller ($p = 0.0077$) than that of CC (0.1277).

In adaptive TT, the or-combination provides the lowest cost (i.e., highest effectiveness). The cost of the or-combination is 10% smaller than that of the and-combination (0.0461 vs. 0.0513, respectively); however, it is not statistically significantly smaller ($p = 0.1121$). The percentage cost reduction provided by the or-combination with respect to CC (−34%: 0.0461 vs. 0.0699) and cosine (−18%: 0.0461 vs. 0.0563) is noticeable. Although these are impressive results, none of the adaptive test results is statistically significantly smaller than the rest of the adaptive test observations. Only the cost of the “and-combination” (0.0513) is close to being statistically significantly smaller ($p = 0.0975$) than that of CC (0.0699).

Note that in adaptive TT, in terms of the best combination method, we observe a contradiction between training and test results: During training, the and-combination outperformed the or-combination; during testing, the reverse was true. This can be explained by the fact that what we observed during training is a not “strongly” statistically significant ($p = 0.0902$; i.e., the p value is relatively large and somewhat strongly significant; i.e., we have suggestive evidence). As indicated earlier, we have a similar observation for the comparison of the and- and or-combination test results ($p = 0.1121$). So, it can be arguably claimed that the test observations are in agreement with those of training.

The adaptive tracking C_{Det} values are statistically significantly smaller than those of the corresponding static observations, and p values are as follows: CC: 0.0555, almost 0.05; cosine: 0.0020; and-combination: 0.0014; or-combination: 0.0019. The TT results (especially the adaptive ones) are good enough for practical settings.

Further Discussion

In this section, we briefly describe our experimental observations associated with some other approaches. There is a large room of improvement in NED, therefore, we further experimented with NED. While performing these experiments, we used (a) word stopping and the longest stoplist, and (b) the incremental *idf* approach for the selection of the document indexing terms.

Use of Other Similarity Measures for NED

For understanding the influence of similarity measures on NED, we also experimented with five additional similarity measures: Dice, Hellinger, Jaccard, Okapi, and overlap. These measures are commonly used in TDT or IR. For example, for TDT applications, Luo et al. (2007) used

TABLE 12. New event detection (NED) standalone use of similarity measures training and test effectiveness (Best test-C_{Det} value is in bold).

Similarity measure	Stemmer	Vector size (*dn*)	Training Min. C_{Det}	Test C_{Det} (relative +% change)[a]
CC	F6	60	0.5599	0.6359 (28.5)
Cosine	LM	All terms	0.5777	**0.4948** (0.0)
Dice	LM	All terms	0.5669	0.5165 (10.4)
Hellinger	LM	All terms	0.6207	0.5638 (14.0)
Jaccard	LM	All terms	0.5664	0.5154 (4.2)
Okapi	LM	50	0.5424	0.5249 (2.1)
Overlap	F5	30	0.6573	0.7700 (55.6)

CC = cover coefficient.

[a] Relative percentage change with respect to C_{Det} value of cosine.

Okapi; Franz, McCarley, Ward, and Zhu (2001) used a symmetrized version of the Okapi formula; and Brants, Chen, and Farahat (2003) used the Hellinger measure. The definitions of the Dice, Jaccard, and overlap similarity measures can be seen in Anderberg (1973) and in Salton (1989). In the experiments, we use *tf.idf* term weights for term selection. We also use these weights in the Dice, Jaccard, and overlap similarity measures for term weighting. With these five additional similarity measures, the window size is taken as 12 days. We experimented with all stemmers (NS, F5, F6, and LM) and all document vector lengths used in the previous experiments (see Table 5). Full details of the training experiments with these similarity measures can be seen in Bağlıoğlu (2009, pp. 71–72) [He also reported another version of Okapi that uses chronological term ranking (Troy & Zhang, 2007), which is outside the scope of this study.] We already reported the CC and cosine results in the previous sections, but they are also reported here for comparison.

The results reported in Table 12 show that the cosine similarity measure provides the best performance (the lowest C_{Det} value of 0.4948) during testing. If the cosine observation is taken as a reference point, the percentage reduced costs provided by all other measures are positive (Rather than reducing, they indeed increase the cost with respect to that of cosine.) Okapi, with the C_{Det} value of 0.5249 and the percentage cost increase of 2.1%, provided the second-best performance after cosine. The Jaccard measure introduced a cost increase of 4.2%. The others introduced a cost increase between 10.4 and 55.6%. In five of seven cases (viz., cosine, Dice, Hellinger, Jaccard, and Okapi measures), the best performance was observed with the lemmatizer-based stemmer LM; however, in two cases (CC and overlap similarity measures), the simple term truncation approaches (F5, F6) outperformed LM. In four cases of seven, it was better to use all terms of news stories for indexing. The use of cosine yielded statistically significantly lower C_{DET} values than two of the other similarity measures, CC and overlap (for both cases, $p < 0.05$).

We also experimented with all possible paired and- and or-combinations of these similarity measures. In the case of the and-combination, the highest cost reduction was observed with the Okapi–overlap combination. It lowered the cost about −5% (C_{Det} value of 0.5012 vs. the best of the Okapi and overlap C_{Det} values: 0.5249) when they were used alone. The other cases of the and-combination provided either small improvements or decreases in effectiveness. In the case of the or-combination, the lowest C_{Det} value (0.4550) is observed with the CC–cosine combination. For this case, the percentage reduced cost is −8.04% (0.4550 vs. the best of the CC and cosine C_{Det} value when they are used alone: 0.4948). The Okapi–cosine or-combination provides the second-best C_{Det} value of 0.4717, and a percentage reduced cost of −4.67%. The other cases of and- and or-combinations provided a small improvement or decrease in the performance. More information about test results can be seen in Kardaş (2009, pp. 52–56). The results indicate that combining the results of different similarity measures improves the performance in some cases. The CC–cosine or-combination that gives the C_{Det} value of 0.4550 (with −8.04% percentage cost reduction) is noticeable; however, none of the combination results is statistically significantly smaller than those of the stand-alone cases used for the combination.

Our test results suggest that in terms of effectiveness, NED in Turkish is similar to other language cases. For example, Zhang, Zi, and Wu (2007, their Table 4) provided some state-of-the-art observations for the TDT3 corpora that include documents in English and Mandarin. Their NED normalized C_{Det} values for five different approaches were between "0.5413 and 0.6493" (average value = 0.5973). Our test phase C_{Det} values (Table 12) for the stand-alone use of the seven similarity measures were between "0.4948 and 0.7700" (average value = 0.5745). Allan et al. (2003, their Figure 3) also provided similar scores for the Hindi language.

Use of Named Entities

We also experimented with the use of named entities. For this purpose, we created a named entity collection of 60,267 items (Uyar, 2009). In this collection, the human names table was generated by using the Web site of "Türk Dil Kurumu"—Turkish Language Association (TDK). The TDK Web site (http://www.tdk.gov.tr) provides a dictionary of person names. The personnel, student, and high-school-student information databases of Bilkent University also were used for human names. For location names, address records of personnel and student information databases of Bilkent University were used; city (in Turkish "şehir"), county (ilçe), and district (semt) names also are inserted into the database. An organization names table was manually created by using frequently used organization names such as TRT, TÜBİTAK, MEB, and so on. In addition to this list, we also used some intuitive ways of recognizing named entities by looking at the contexts of words that begin with capital letters.

In the experiments, we used three different similarity scores using document vectors based on (a) words other than named entities, (b) only named entities, (c) all words, and (d) the triangularization approach described by Kumaran et al. (2004). In our experiments, among these four approaches, the case that uses all words alone provided the best performance. The reason of our mediocre results with the named entities can be attributed to the possibility that our test-collection topic stories are not conducive to the use of named entities. Kumaran and Allan (2004) also had similar observations.

Use of Decaying Function for Decreasing Influence of Earlier Stories

We also used a time decaying–weight function method that assigned lower importance to earlier stories in confidence score calculations as defined by Yang et al. (1998). In this method, as a document becomes older, its influence on decisions becomes smaller. In the experiments, we only considered the CC and cosine similarity measures.

First, we used the time-window sizes that we already had chosen for these measures (CC: 12 days, cosine: 14 days). In the experiments, we observed that this approach decreases the performance. We speculate that since we use a properly chosen window size, the time decay is unnecessary, and that if used, it has a degenerative effect (Kardaş, 2009, pp. 57–58). However, if the average time span of typical events cannot be predicted correctly, then in that case, the time decay approach might have a positive influence. To confirm this intuition, we performed additional experiments and gradually increased the window size. In these experiments, for CC beginning with about a time-window size of 50 days, we started to get a performance that was not as good as a window size of 14 days, but it was close to it. Finally, using a considerably larger window size of 100 days, the effectiveness level reached the level of 14 days that uses no time decay. However, for larger time-window sizes, as early as 120 days, system effectiveness starts to decrease again. For the cosine measure, we had similar observations, but for a different numbers of days (e.g., with a time-window size of 21 days, it gives a good performance, and then suddenly worsens thereafter). The negative impact of larger window sizes can be attributed to the fact that when the window is too wide, some old events are still able to sustain their influence on the system. In a real-time environment, the use of large time-window sizes can be detrimental to efficiency. The computational cost can be considerable if stories come with a small time delay (Luo et al., 2007).

Conclusions and Future Work

The multi-resource, large-scale test collection BilCol-2005 that we constructed in this study is a significant contribution to the evaluation resources available for TDT applications. Using BilCol-2005, we provide pioneering benchmark observations for NED and TT in Turkish. We show that TT in Turkish has a performance similar to those in other languages and can be used in operational settings (Can et al., 2008a; Öcalan, 2009). Our NED results also are comparable to those in other languages; however, for NED, there is still much room for improvement (Allan et al., 2003; Allan et al., 2000).

For similar TDT applications in Turkish, we recommend using

- The cosine similarity measure: Furthermore, system effectiveness with the cosine measure can be improved if its results and those of CC are combined using the or-combination method. This approach in NED provides a major decrease in MRs, which can be important in intelligence applications.
- A lemmatizer-based stemmer for indexing: With the cosine similarity measure, it provides a statistically significantly better NED effectiveness than those of no stemming and the fixed prefix method; if some decrease in effectiveness is tolerable and a quick implementation is needed, then the fixed prefix stemming method can be used; indexing with no stemming is not an acceptable option.
- A broad stoplist: The use of stoplist has a statistically significant impact on NED effectiveness, and as the list size increases, effectiveness improves; however, the gain in effectiveness diminishes as the size increases.
- Adaptive tracking: It provides a statistically significantly higher effectiveness than all corresponding static tracking methods.

There are several future research possibilities. The NED problem is difficult, and further research is needed. The use of named entities in NED needs further investigation. In practical settings, it is important to display events that may attract common news-consumer interests (Allan et al., 2005). It would be interesting to know whether the utility of the window size or decay function showed any systematic variation with topic type. A fuzzy logic approach can be developed for combining the confidence scores of different similarity measures.

Acknowledgments

We thank our colleagues, friends, and students for their topic annotations. We are grateful to the anonymous referees for their comments; their pointers have significantly improved the article. We also thank Sengor Altingovde and Jon Patton for their helpful comments. This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under Grant 106E014. Any opinions, findings, and conclusions or recommendations expressed in this article belong to the authors and do not necessarily reflect those of the sponsor; therefore, no official endorsement should be inferred.

References

Allan, J. (Ed.). (2002a). Topic detection and tracking: Event-based information organization. Norwell, MA: Kluwer Academic.

Allan, J. (2002b). Introduction to topic detection and tracking. In J. Allan (Ed.), Topic detection and tracking: Event-based information organization (pp. 1–16). Norwell, MA: Kluwer Academic.

Allan, J., Harding, S.M., 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, Track 4 (p. 101.1).

Allan, J., Lavrenko, V., & Connell, M.E. (2003). A month to topic detection and tracking in Hindi. ACM Transactions on Asian Language Information Processing, 2(2), 85–100.

Allan, J., Lavrenko, V., & Jin, H. (2000). First story detection in TDT is hard. In Proceedings of the Ninth International Conference on Information and Knowledge Management (pp. 374–381). New York: ACM Press.

Allan, J., Lavrenko, V., & Swan, R. (2002). Explorations within topic tracking and detection. In J. Allan (Ed.), Topic detection and tracking: Event-based information organization (pp. 197–224). Norwell, MA: Kluwer Academic.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 37–45). New York: ACM Press.

Altingovde, I.S., Demir, E., Can, F., & Ulusoy, O. (2008). Incremental cluster-based retrieval using compressed cluster-skipping inverted files. ACM Transactions on Information Systems, 26(3).

Amini, L. (2007). Stream processing: What's in it for you? IBM T.J. Watson Research Center. Retrieved March 8, 2008, from http://www-05.ibm.com/nl/events/presentations/stream_processing_whats_in_ it_for_ you.pdf

Anderberg, M.R. (1973). Cluster analysis for applications. New York: Academic Press.

Bağlıoğlu, Ö. (2009). New event detection using chronological term ranking. Unpublished master's thesis, Bilkent University, Computer Engineering Department. Retrieved June 21, 2009, from http://www.cs.bilkent.edu.tr/∼canf/bilir_web/theses/ozgurBagliogluThesis.pdf[1]

BilCol-2005. (2009). Bilkent TDT collection for the year 2005. Retrieved July 10, 2009, from http://www.cs.bilkent.edu.tr/∼canf/bilcol/bilcol.html

Brants, T., Chen, F., & Farahat, A. (2003). A system for new event detection.In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 330–337). New York: ACM Press.

Can, F. (1993). Incremental clustering for dynamic information processing. ACM Transactions on Information Systems, 11(2), 143–164.

Can, F., Altingovde, I.S., & Demir, E. (2004). Efficiency and effectiveness of query processing in cluster-based retrieval. Information Systems, 29(8), 697–717.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., & Uyar, E. (2008a). Bilkent News Portal: A personalizable system with new event detection and tracking capabilities. In Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval (p. 885). New York: ACM Press.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, C., & Vursavas, O.M. (2008b). Information retrieval on Turkish texts. Journal of the American Society for Information Science and Technology, 59(3), 407–421.

Can, F., & Ozkarahan, E.A. (1990). Concepts and effectiveness of the cover coefficient-based clustering methodology for text databases. ACM Transactions on Database Systems, 15(4), 483–517.

Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K., & Liberman, M. (2002). Corpora for topic detection and tracking. In J. Allan (Ed.), Topic detection and tracking: Event-based information organization (pp. 33–66). Norwell, MA: Kluwer Academic.

Connell, M., Feng, A., Kumaran, G., Raghavan, H., Shah, C., & Allan, J. (2004). UMass at TDT 2004. Retrieved July 30, 2008, from http://maroo.cs.umass.edu/pub/web/getpdf.php?id=507

Croft, B., Metzler, D., & Strohman, T. (2009). Search engines: Information retrieval in practice. Boston: Addison-Wesley.

Dolamic, L., & Savoy, J. (2010). When stopword lists make the difference. Journal of the American Society for Information Science and Technology, 61(1), 200–203.

Elsayed, T., & Oard, D.W. (2005). On evaluation of adaptive topic tracking systems. In Proceedings of the 28rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 597–598). Salvador, Bahia, Brazil: ACM.

Fiscus, J.G., & Doddington, G.R. (2002). Topic detection and tracking evaluation overview. In J. Allan (Ed.), Topic detection and tracking: Event-based information organization (pp. 17–31). Norwell, MA: Kluwer Academic.

Fiscus, J., & Wheatley, B. (2004). Overview of the TDT 2004 evaluation and results. Retrieved November 29, 2009, from www.itlnist.gov/iad/mig/tests/tdt/2004/papers/NIST-TDT2004.ppt

Franz, M., McCarley, J.S., Ward, T., & Zhu, W.-J. (2001). Unsupervised and supervised clustering for topic tracking. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 310–317). New York: ACM Press.

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 224–231). New York: ACM Press.

He, D., Brusilovsky, P., Ahn, J., Grady, J., Farzan, R., Peng, Y., et al. (2008). An evaluation of adaptive filtering in the context of realistic task-based information exploration. Information Processing & Management, 44(2), 511–533.

Hodges, J.L., & Lehmann, E.L. (1964). Basic concepts of probability and statistics. San Fransisco: Holden-Day.

Kardaş, S. (2009). New event detection and tracking in Turkish. Unpublished master's thesis, Bilkent University, Computer Engineering Department. Retrieved June 21, 2009, from http://www.cs.bilkent.edu.tr/∼canf/bilir_web/theses/suleymanKardasThesis.pdf

Kowalski, G.J. (1997). Information storage and retrieval systems: Theory and implementation (2nd ed.). Boston: Kluwer Academic.

Kumaran, G., & Allan, J. (2004). Text classification and named entities for new event detection. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 297–304). New York: ACM Press.

Kumaran, G., Allan, J., & McCallum, A. (2004). Classification models for new event detection. CIIR Tech. Rep. No. IR-362. Retrieved December 2, 2009, from http://ciir.cs.umass.edu/pubfiles/ir-362.pdf

Kurt, H. (2001). On-line new event detection and tracking in a multi-resource environment. Unpublished master's thesis, Bilkent University, Computer Engineering Department. Retrieved December 20, 2005, from http://www.cs.bilkent.edu.tr/tech-reports/2001/BU-CE-0110.ps.gz

Leek, T., Schwartz, R., & Sista, S. (2002). Probabilistic approaches to topic detection and tracking. In J. Allan (Ed.), Topic detection and tracking: Event-based information organization (pp. 67–83). Norwell, MA: Kluwer Academic.

Luo, G., Tang, C., & Yu, P.S. (2007). Resource-adaptive new event detection. In Proceedings of the 27th International Conference on Management of Data (pp. 497–508). New York: ACM Press.

Manmatha, R., Feng, A., & Allan, J. (2002). A critical examination of TDT's cost function. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 403–404). New York: ACM Press.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In Proceedings of Eurospeech (Vol. 4, pp. 1895–1898).

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., et al. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In Proceedings of the Human Language Technology Conference, (pp. 280–285). San Francisco: Morgan Kaufmann.

NIST. (2008). NIST tools. Retrieved June 22, 2008, from http://www.nist.gov/speech/tools/

Nuray, R., & Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. Information Processing & Management, 42(3), 595–614.

Öcalan, H.Ç. (2009). Bilkent News Portal: A system with new event detection and tracking capabilities. Unpublished master's thesis, Bilkent

[1] All Bilkent graduate theses also are accessible from the search facility of the Bilkent University Library: http://library.bilkent.edu.tr/ (A time delay is possible for database update).

University, Computer Engineering Department. Retrieved June 21, 2009, from http://www.cs.bilkent.edu.tr/∼canf/bilir_web/theses/cagdasOcalanThesis.pdf

Papka, R. (1999). On-line new event detection, clustering, and tracking. Unpublished doctoral dissertation, University of Massachusetts at Amherst, Computer Science Department.

Pouliquen, B., Steinberger, R., Ignat, C., Kasper, E., & Temnikova, I. (2004). Multilingual and cross-lingual news topic tracking. In Proceedings of the 20th International Conference on Computational Linguistics (ACL) (Article no. 959). Morristown, NJ: Association for Computational Linguistics.

Radev, D., Otterbacher, J., Winkel, A., & Balir-Goldensohn, S. (2005). News InEssence: Summarizing online news topics. Communications of the ACM, 48(10), 95–98.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of ınformation by computer. Reading, MA: Addison-Wesley.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Stokes, N., & Carthy, J. (2001). Combining semantic and syntactic document classifiers to improve first story detection. Poster presented at the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 424–425). New Orleans, LA: ACM.

TDT. (2002). The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan. Tech. Rep. Version 1.1, National Institute of Standards and Technology. Retrieved December 2, 2009, from http://www.itl.nist.gov/iad/mig/tests/tdt/2002/evalplan.html

TDT. (2004). Annotation manual: Version 1.2. Retrieved January 9, 2007, from http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf

TDT. (2008). Topic detection and tracking evaluation. Retrieved June 18, 2008, from http://www.itl.nist.gov/iaui/894.01/tests/tdt/

Troy, A.D., & Zhang, G.-Q. (2007). Enhancing relevance scoring with chronological term rank. In Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 599–606). New York: ACM Press.

Uyar, E. (2009). Near-duplicate news detection using named entities. Unpublished master's thesis, Bilkent University, Computer Engineering Department. Retrieved June 21, 2009, from http://www.cs.bilkent.edu.tr/∼canf/bilir_web/theses/erkanUyarThesis.pdf

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

Vural, A. (2002). On-line new event detection and clustering using the concepts of the cover coefficient-based clustering methodology. Unpublished master's thesis, Bilkent University, Computer Engineering Department. Retrieved July 6, 2008, from http://www.cs.bilkent.edu.tr/tech-reports/2002/BU-CE-0218.pdf

Yang, Y., Carbonell, J., Brown, R., Lafferty, J., Pierce, T., & Ault, T. (2002). Multi-strategy learning for topic detection and tracking. In J. Allan (Ed.), Topic detection and tracking: Event-based information organizationbreak (pp. 85–114). Norwell, MA: Kluwer Academic.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., & Liu, X. (1999). Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4), 32–43.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and on-line event detection. In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 28–36). New York: ACM Press.

Yi, X., & Allan, J. (2008). Evaluating topic models for information retrieval. In Proceedings of the 17th Conference on Information and Knowledge Management (pp. 1431–1432). New York: ACM Press.

Yu, C.T., & Meng, W. (1998). Principles of database query processing for advanced applications. San Francisco: Morgan Kaufmann.

Zhang, K., Zi, J., & Wu, L.G. (2007). New event detection based on indexing-tree and named entity. In Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 215–222). New York: ACM Press.