New Event Detection and Topic Tracking in Turkish

FAZLI CAN¹, SEYIT KOCBERBER, OZGUR BAGLIOGLU
SULEYMAN KARDAS, H. CAGDAS OCALAN, ERKAN UYAR

Bilkent Information Retrieval Group
Computer Engineering Department
Bilkent University
Ankara, Turkey 06800

This is a preprint of an article accepted for publication in Journal of the American Society for Information Science and Technology copyright © [2009] (American Society for Information Science and Technology) October 1, 2009

Abstract

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 (IR), 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 is also investigated. We show that it is possible to deploy TT applications in Turkish that can be used in operational settings.

Keywords: New event detection (NED), news portal, topic detection and tracking (TDT), TDT test collection construction, Turkish.

INTRODUCTION

Information explosion has new dimensions with the advances in information technologies. For example, the number of news resources on the Web has exponentially increased in the last decade. Multi-resource 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. (In the rest of the paper, 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,
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.” 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.” 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. It should be noted 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 like "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 as follows.

- New Event Detection (NED): aims to recognize the first story for a new event that had not been discussed before. This problem is also 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 to 4) that are 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 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 is constructed by 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 paper 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 paper 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.
New Event Detection and Topic Tracking in Turkish

(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.

![Sliding time-window](image)

**FIG. 1. Sliding time-window.**

(Different shapes represent different events.)

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 (dissimilar). “Unique” (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, Chapter 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 (Yang, Pierce, & Carbonell, 1998; Papka, 1999). Here the assumption is that the stories related to an event are near to each other in terms of their occurrence time. In this study, we also use the sliding time-window concept for NED.

Clustering concepts are used in various TDT studies. Yang, Pierce, & Carbonell (1998) use hierarchical and non-hierarchical document clustering algorithms to solve the NED problem. In their approach, they pay attention to temporal information and use 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) employs link-based clustering concepts in various TDT tasks. Allan, Lavrenko, & Jin (2000) study the difficulties of finding new events with the traditional single-pass clustering approach, and show that with certain assumptions NED effectiveness can be predicted from that of TT, since these two tasks are related. In this paper, our NED method, which is based on the time-window concept, is inspired by the incremental clustering strategy and similar to the ones available in the literature.

Combining the results of different approaches is studied by various research groups. Hatzivassiloglou, Gravano, & Maganti (2000) study the problem of combining the results of different similarity functions and propose a theoretically justified statistical model that performs as good as or better than that of an exhaustive search. Stokes & Carty (2001) use a composite document representation that involves concept representation based on lexical chains. Yang, Carbonell, Brown, Lafferty, Pierce, & Ault (2002) study a combination system called BORG (Best Overall Results Generator for tracking) by using the results of various classifiers and examining their decision error trade-off (DET) curves. Kumaran & Allan (2004) examine the effects of stopwords and the combination of different document vectors.
(named entity vectors, non-named entity vectors) on NED. In this work, we examine the effects of using named entities, wordstopping, 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, Brusilovsky, Ahn, Grady, Farzan, Peng, Yang, & Rogati, 2008). Luo et al. (2007) study a practical new event detection system using IBM’s Stream Processing Core middleware (Amini, 2007). The work of Allan, Harding, Fisher, Bolivar, Guzman-Lara, & Amstutz (2005) present 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, Kocheberber, Baglioglu, Kardas, Ocalan, Uyar, 2008a; Kardaş, 2009).

There is limited research on topic detection and tracking 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 (2002) conducted the only TDT study for Turkish other than ours using 46,530 stories and 15 annotated events. Our communication with Kurt reveals that his test collection has been misplaced and is unavailable for further research.

**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 (Cagdas, 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), 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 way. 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.

<table>
<thead>
<tr>
<th>News Source</th>
<th>No. of News Stories</th>
<th>Percent of All Stories</th>
<th>Download Amount (MB)</th>
<th>Net Amount (MB)</th>
<th>Avg. No. of Words per Story</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN Türk</td>
<td>23,644</td>
<td>11.3</td>
<td>1,008.3</td>
<td>66.8</td>
<td>271</td>
</tr>
<tr>
<td>Haber 7</td>
<td>51,908</td>
<td>24.8</td>
<td>3,629.5</td>
<td>107.9</td>
<td>238</td>
</tr>
<tr>
<td>Milliyet Gazetesi</td>
<td>72,233</td>
<td>34.5</td>
<td>508.3</td>
<td>122.5</td>
<td>218</td>
</tr>
<tr>
<td>TRT</td>
<td>18,990</td>
<td>9.1</td>
<td>937.9</td>
<td>18.3</td>
<td>121</td>
</tr>
<tr>
<td>Zaman Gazetesi</td>
<td>42,530</td>
<td>20.3</td>
<td>45.3</td>
<td>33.7</td>
<td>97</td>
</tr>
<tr>
<td>All together</td>
<td>209,305</td>
<td>100.0</td>
<td>6,129.3</td>
<td>349.2</td>
<td>196*</td>
</tr>
</tbody>
</table>

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

**Topic Profiles**

For each topic they select annotators are required to complete a topic profile (TDT, 2004). This process aims to provide documentation for the topics being annotated and help
annotators to properly investigate the related seminal event. The profiles are also used during the annotation process (see below). 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),
- **Topic type:** defined below.

Like the ones used in the TDT research initiative (Cieri et al., 2002; TDT, 2004) there are 13 topic types: 1) elections, 2) scandals/hearings, 3) legal/criminal cases, 4) natural disasters, 5) accidents, 6) acts of violence or war, 7) science and discovery news, 8) financial news, 9) new laws, 10) sports news, 11) political and diplomatic meetings, 12) celebrity/human interest news, and 13) 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, the annotators were allowed to see each other’s 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 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 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 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. In order
to urge annotators to see the whole document contents 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 below.

**Step-1 (Search with the seed):** ETracker searches the collection for tracking stories by using the seed (first) story as a query. The annotator decides if the results are on topic or not, 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/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 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.

**Step-2 (Search with the profile information):** ETracker ranks by using the profile description words as a query.

**Step-3 (Search with on-topic stories):** ETracker takes the first three on-topic stories of step-1 and step-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.

**Step-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. In order 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 minutes. 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 ½. This means that for example, if an annotator finds 10 on-topic stories in their list, they must also 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.

**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 this table for the “Onur Air’in Avrupa’dan 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, and 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; http://www.cs.bilkent.edu.tr/~canf/BilCol/BilCol.html)

### TABLE 2. News categories and number of annotated topics in each category.

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

* Due to double category assignment to 6 topics there are total of 86 topics.

### TABLE 3. Information about sample topics and some averages for BilCol2005.

<table>
<thead>
<tr>
<th>Sample Topic (Topic No. in BilCol-2005)</th>
<th>No. of Tracking Stories</th>
<th>Life Span (day)</th>
<th>fn=100</th>
<th>fn=50</th>
<th>fn=25</th>
<th>fn=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onur Air’in Avrupa’da yasaklanması – Banning of Onur Air in Europe (2)</td>
<td>159</td>
<td>203</td>
<td>154</td>
<td>154</td>
<td>148</td>
<td>105</td>
</tr>
<tr>
<td>Londra metrosunda patlama – Explosion in London subway (6)</td>
<td>454</td>
<td>175</td>
<td>440</td>
<td>419</td>
<td>376</td>
<td>236</td>
</tr>
<tr>
<td>400 koyun intihar etti – 400 sheep commit suicide (10)</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Mortgage Türkiye’de – Mortgage has arrived in Turkey (14)</td>
<td>375</td>
<td>357</td>
<td>60</td>
<td>41</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>Attilâ İlhan vefat etti – Attilâ İlhan passed away (21)</td>
<td>40</td>
<td>70</td>
<td>40</td>
<td>37</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>Sahte rakı – Counterfeit rakı (48)</td>
<td>323</td>
<td>182</td>
<td>316</td>
<td>291</td>
<td>255</td>
<td>197</td>
</tr>
<tr>
<td>İlk yüz nakli – First face transplantation (61)</td>
<td>14</td>
<td>17</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>10</td>
</tr>
</tbody>
</table>

Averages for all 80 topics of BilCol-2005

<table>
<thead>
<tr>
<th></th>
<th>No. of stories on first fn days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fn=100</td>
</tr>
<tr>
<td>Averages</td>
<td>73</td>
</tr>
</tbody>
</table>

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

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. All together, there were 39 native speaker annotators. They used interpretation rules that are similar to those used in the TDT
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 the average, there are 73 (median: 32, minimum: 5, maximum: 454) tracking stories for each topic. On the average, annotators spent, not counting breaks from work, 109 (median: 80, minimum: 20, maximum: 825) minutes 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). The definitions of these effectiveness measures are as follows.

- False Alarm (FA) = number of tracking stories labeled as new event / total number of tracking stories.
- Miss Rate (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 0.

FA and MR are shown with a curve by using false alarms and miss rates 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 ROC (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, e.g. it expands the high performance region, for more information please refer to (Fiscus, Doddington, 2002, p. 24). DET curves provide a visualization of the tradeoff 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 more preferable than story-weighted estimates (Doddington, 1998, Fiscus & Doddington, 2002, p. 22). Numerical examples for story- and topic-weighted approaches can be seen in (Bagloğlu, 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 a false alarm, and they are pre-specified,
- $P_{target} = 0.02$, 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 pre-specified numerical values given above are consistently used in TDT performance evaluation (Fiscus & Doddington, 2002; Fiscus & Wheatley, 2004, Manmatha, Feng, Allan, 2002).

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

$$\left( C_{\text{Det}} \right)_{\text{Norm}} = \frac{C_{\text{Det}}}{\text{Minimum} \left\{ C_{\text{Miss}} \cdot P_{\text{targets}} \cdot C_{\text{FA}} \cdot (1 - P_{\text{target}}) \right\}}$$

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 $C_{\text{Det}}$ 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 $\theta_{\text{on}}$ are declared as new, other cases are treated as old. At each threshold value, the misses and false alarms 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, and McCallum 2004). Different NED systems are compared based on their minimum cost. This minimum cost is defined as $\text{Min. } C_{\text{Det}} = \min \left\{ \left( C_{\text{Det}} \right)_{\text{Norm}} \right\}$ where $\left( C_{\text{Det}} \right)_{\text{Norm}} \in S$, 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’ $C_{\text{Det}}$ value using an alpha level of 0.05 for the significance 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 $C_{\text{Det}}$ values, which were identified as potential outliers, were eliminated. A potential outlier is defined as an observation that is more than 2.5 standard deviations above or below the mean. In other words, we first compute the differences for each pair and we extract the pair whose value is more or less than 2.5 standard deviations from the mean. This approach is repeated with the remaining observations. We observed only one and two outliers 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 and if the newcomer is different (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 and if it is similar enough it is assumes 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, and $dn$ is the number of non-zero items in the document vector ($dn << n$). 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 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,

\[
w(t, d) = (1 + \log tf(t, d)) \cdot \log \left( \frac{N_i}{n_i} \right)
\]

In this formula, \(w(t, d)\) is the weight of term \(t\) in document \(d\); \(tf(t, d)\) is the number of occurrences of term \(t\) in document \(d\); \(\log \left( \frac{N_i}{n_i} \right)\) is the \(idf\) (inverse document frequency) component of the formula, where \(n_i\) is the number of stories in the collection that contains one or more occurrence of term \(t\) including the newcomer, and \(N_i\) is the number of accumulated stories so far in the collection. Hence, \(n_i\) and \(N_i\) (and therefore the \(idf\) values) are incrementally computed. A similar approach is used in other studies (e.g., Yang et al., 1998). We use an auxiliary corpus containing the 2001-2004 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 weighting method is used with the cosine similarity measure. In a later section the same term weighting formula is also used with the Dice, Jaccard, and 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 (van Rijsbergen, 1979; Salton, 1989; Salton & Buckley, 1988).

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

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

**Cover Coefficient-based Similarity Measure.** According to the cover coefficient (CC) concept, the coverage (defined below) of \(d_i\) by \(d_j\), based on a document by term D matrix of dimensions \(m\) and \(n\), is expressed as follows (Can & Ozkarahan, 1990; Can, 1993).

\[
c_{ij} = \sum_{k=1}^{n} \left[ \alpha_i \cdot d_{ik} \right] \left[ \beta_k \cdot d_{jk} \right], \text{ where } \alpha_i = \left[ \sum_{k=1}^{n} d_{ik} \right]^{-1}, \beta_k = \left[ \sum_{i=1}^{m} d_{ik} \right]^{-1}, 1 \leq i \leq m, 1 \leq j \leq m
\]

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.); \(\alpha_i\) and \(\beta_j\) are, respectively, 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), and then 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; \(c_{ij}\) 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

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 \( d_2 \) 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” (\( d_l \)) 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.

[FIG. 3. From left to right: example binary D matrix \( (m=5, n=4) \), hierarchical representation of the two-stage probability model for \( d_1 \) of the D matrix, and C: cover coefficient matrix (some values are approximate).]

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 \( \beta \) 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 \( \beta \) values, and c) using a variant of the incremental approach by modifying the values \( \alpha_i \) and \( \beta_k \) as follows for smoothing their effects on \( c_{ij} \).

\[
\alpha_i = \left[ \log \sum_{j=1}^{n} d_{ij} \right]^{-1}, \quad \beta_k = \left[ \log \sum_{i=1}^{m} d_{ik} \right]^{-1}
\]

In this approach, we take logarithms in order to not overly diminish the \( d_{ik} \) and \( d_{jk} \) values in the \( c_{ij} \) formula, since the original \( \alpha_i \) and especially \( \beta_k \) 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, p. 104, p. 116). In the experiments, the best results with CC are obtained with the logarithmic approach and in this paper 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 and can be seen in (Vural, 2002).

**New Event Detection (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).
**Standalone 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 pre-determined threshold, where the confidence score of \(d\) is defined as \(\max_{d_{i \in \text{window}}} \left( \text{sim}(d, d_{i}) \right)\) i.e., the maximum similarity observed between \(d\) and the members of the time-window. The NED method is defined in Figure 4.

1. Obtain the terms of \(d\).
2. Update the \(idf\) statistics using the terms of \(d\).
3. Select the highest \(tfdidf\) 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 \(d\), \(\max_{d_{i \in \text{window}}} \left( \text{sim}(d, d_{i}) \right)\) i.e., the maximum similarity value among the computed values.
5. If the confidence score is less than the NED threshold \(\theta_{\text{on}}\) (\(\theta_{\text{CC}}\) for CC and \(\theta_{\text{cosine}}\) for cosine), then \(d\) is labeled as the first story of a new event.

![FIG. 4. New event detection (NED) - cosine and CC- methods.](image)

The pre-determined thresholds for the cosine and the CC measures, namely \(\theta_{\text{cosine}}\) and \(\theta_{\text{CC}}\), are obtained by training; these values provide the Min. C Det values for the respective measures.

**Combination Methods: Experimental Evidence for Our Intuition.** In this paper, we hypothesize that decisions of two different similarity measures can be combined to achieve effectiveness better than standalone use. 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 data set which is 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 data set for decision-making. It shows that these two measures agree on 23 decisions jointly, seven cases are only detected by cosine, six cases are only detected by CC; all together we have 36 correct decisions out 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 false alarms generated by tracking news stories during NED. It shows that the cosine and CC measures, respectively, incorrectly detect 110 and 127 tracking stories as new events and among these they have 50 common false drops. Their agreement in making wrong decisions together is about 42%.

![FIG. 5. NED consistency of CC and cosine decisions during training for 50 events: left Venn diagram is for true matches, right Venn diagram is for false alarms.](image)

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: 1) 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). 2) 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 false alarm 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 number 5 of Figure 4 as follows by using one of the combination alternatives given below.

- **And-combination:** if \( x < \theta_{\text{cosine}} \) and \( y < \theta_{\text{CC}} \), then \( d \) is labeled as the first story of a new event.
- **Or-combination:** if \( x < \theta_{\text{cosine}} \) or \( y < \theta_{\text{CC}} \), 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, i.e.,

\[
x = \max_{d_k \in \text{window}} \left( \text{sim}_{\text{cosine}}(d, d_k) \right) \quad \text{and} \quad y = \max_{d_k \in \text{window}} \left( \text{sim}_{\text{CC}}(d, d_k) \right).
\]

**Topic Tracking (TT) Methods**

The TT task uses (usually 1 to 4) a few number of 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).

**Standalone 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 & 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 \( \theta_{\text{adapt}} \) (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 \( \theta_{\text{adapt}} \) 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 (van Rijsbergen, 1979; Altingovde, Demir, Can, Ulusoy, 2008). 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 at the same time follow topic changes by focusing on new developments.

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 confidence score: \( \text{sim}(T, d) \), the similarity between the current story \( d \) and the target topic \( T \).
5. If the confidence score is above the TT threshold \( \theta_{\text{on}} \) \( (\theta_{\text{CC}} \text{ for CC and } \theta_{\text{cosine}} \text{ for cosine, obtained by training}) \), 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 \( \theta_{\text{on}} \) is above the pre-determined adaptation threshold \( \theta_{\text{adapt}} \) (different \( \theta_{\text{adapt}} \) 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 CC- methods.
**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 number 5 of Figure 6 as follows (in the following x and y, respectively, indicate the CC and cosine similarity of the incoming story to the topic under consideration).

- **And-combination:** if \( x > \theta_{\text{cosine}} \) and \( y > \theta_{\text{CC}} \), then the incoming story \( d \) is labeled as a tracking story of \( T \).

- **Or-combination:** if \( x > \theta_{\text{cosine}} \) and \( y > \theta_{\text{CC}} \) 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 eight and last four months, respectively, serve as the training and test collections. The division gives us the opportunity of keeping most of the tracking stories together with their corresponding first stories. For example, dividing the data set into two six-month periods (January to June, and July to December) does not give us that opportunity (see Figure 2). All together, we have 80 topics. For two topics (Bilcol-2005, 2009; topic numbers 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. By this approach all-together in the training and test sets there are 82 topics (i.e., two more than the original 80 topics). Although using “non-overlapping 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 (Xi & Allan, 2008). As can be seen from Table 4, the average number of news stories per topic in the train and test sets are approximately the same 67.16 (3,358/50) and 71.50 (2,288/32) stories.

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Time Span (month.day.year)</th>
<th>No. of Topics</th>
<th>No. of Documents</th>
<th>No. of Tracked Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>01.01.2005 - 08.31.2005</td>
<td>50</td>
<td>141,910</td>
<td>3,358</td>
</tr>
<tr>
<td>Test</td>
<td>09.01.2005 - 12.31.2005</td>
<td>32</td>
<td>67,395</td>
<td>2,288</td>
</tr>
</tbody>
</table>

**Stemming Methods Used in the Study**

In our recent work, it is 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. They are (1) no stemming, so called “austroch algorithm,” (2) first-n, n-prefix, characters of each word – two versions, n=5 and 6-, and (3) a lemmatizer-based stemmer.

- **No-Stemming (NS).** The no-stemming (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 can also be interpreted as a restricted case of character n-grams. In this study, we experiment with F5 and F6, which are experimentally shown to give the best performance in Turkish IR (Can et al., 2008b).
New Event Detection and Topic Tracking in Turkish

- **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 the paper, we prefer the word “stemming” over lemmatization as it is more commonly used.

The fixed prefix approach is a simple method and it 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 is also 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 use the longest stoplist. More detail on stoplist selection is provided in the next section.)

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

With the cosine measure, the combination “all terms and LM” provides the best performance; furthermore, in 15 of the 16 different document vector lengths LM provides the best performance with respect to the other stemmers. For all stemmer cases using all terms gives 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 (Allan, Lavrenko, & Swan, 2002, p. 200).

In terms of the CC measure, F6 with the document vector length of 60 provides the best performance. In fact, F6 yields the best performance (the lowest Min. C_Det value) for most document vector lengths: in 10 cases out of 16 different vector lengths F6 provides 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 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 other stemming approaches. The training and test results for these cases with all stemmers are provided in Table 6 (test results are 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, LM is significantly better than NS ($p < 0.001$), F5 and F6 ($p < 0.05$). When CC is used, F6 is significantly better than NS and F5 ($p < 0.05$), but slightly better than LM ($p < 0.10$). The F6 vs. 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.
TABLE 5. Min. $C_{Det}$ values for CC and cosine measures for NED with various document vector length (nd) and stemmer combinations* (using the longest stopword list).

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

*Best case for each stemmer is in bold and best case of each similarity measure is underlined.

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

<table>
<thead>
<tr>
<th>Stemmer</th>
<th>Cover Coefficient (CC)</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Results</td>
<td>Test Results</td>
</tr>
<tr>
<td></td>
<td>$Min. \ C_{Det}$</td>
<td>Threshold</td>
</tr>
<tr>
<td>NS</td>
<td>0.6411</td>
<td>0.199</td>
</tr>
<tr>
<td>F5</td>
<td>0.5929</td>
<td>0.252</td>
</tr>
<tr>
<td>F6</td>
<td><strong>0.5730</strong></td>
<td>0.254</td>
</tr>
<tr>
<td>LM</td>
<td>0.5982</td>
<td>0.290</td>
</tr>
</tbody>
</table>

In the rest of the paper, 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 stopword list 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 & Allan (2004) show positive influence of category-based word stopping on NED. We investigate the effects of word stopping by using four stopword lists with different sizes. As a baseline, we use a stopword list that contains no words. This is followed by a stopword list that contains 10 frequent words as determined in our recent work on Turkish IR (Can et al., 2008b). We also experiment with two additional semi-automatically generated stoplists containing 147 and 217 words. The larger sets are inclusive of the smaller ones. The stopword list generation approach is defined in (Can et al.,
The first two lists are provided in (Can et al., 2008b) and all of them are available in (Kardaş, 2009).

### TABLE 7. Effects of stopword list size on NED effectiveness with CC and cosine similarity measures.

<table>
<thead>
<tr>
<th>Stopword List Size</th>
<th>Cover Coefficient (CC)</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Results</td>
<td>Test Results</td>
</tr>
<tr>
<td></td>
<td>Min. $C_{Det}$ Threshold</td>
<td>$C_{Det}$</td>
</tr>
<tr>
<td>0</td>
<td>0.7149</td>
<td>0.256</td>
</tr>
<tr>
<td>10</td>
<td>0.6512</td>
<td>0.246</td>
</tr>
<tr>
<td>147</td>
<td>0.5895</td>
<td>0.244</td>
</tr>
<tr>
<td>217</td>
<td>0.5730</td>
<td>0.254</td>
</tr>
</tbody>
</table>

The results presented in Table 7 show that stopword lists and their sizes have an influence on effectiveness. As we increase the stopword list 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 stopword list varies. The training Min. $C_{Det}$ values of the last row of Table 7 can also be seen in Table 5. The experimental conditions that are specified for Table 5 are also 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 stopword list yields Min. $C_{Det}$ and $C_{Det}$ values that are significantly smaller than using no stopwords for both similarity measures ($p < 0.05$). 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 stop list 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 due to stopword use are due to the improper relative weight of $idf$ in the term weighting formula (Dolamic, Savoy, 200). 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 data set (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 paper we use these window sizes for the cosine and CC 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 false alarm values CC provides lower miss rates; for false alarm values greater than about 7 (%) cosine provides lower miss rates.

The optimum training results for each new event detection approach are given in Table 8. The optimum similarity thresholds values are 0.254 and 0.200 for the CC and cosine
measures, respectively. In addition, we observe that optimum thresholds for the and- and or-
combination methods are the same as the case with the standalone use of these similarity
measures. The best performance with the standalone and combined versions are in bold. The
miss rates are much higher than false alarms. When the similarity measures are used alone,
CC provides a slightly smaller cost value (better effectiveness) than that of cosine (0.5599 vs.
0.5777); however, it is not statistically significant. It is observed that in training, 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 and-combination’s (0.6268).

FIG. 7. NED training performance with CC and cosine similarity measures.

TABLE 8. NED training results.
(Best standalone and combination results are in bold, best case is underlined.)

<table>
<thead>
<tr>
<th>Similarity Measure – Combination Method</th>
<th>Miss Rate</th>
<th>False Alarm</th>
<th>Min. (C_{Det})</th>
<th>NED Threshold ((\theta_{on}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.4200</td>
<td>0.0285</td>
<td>0.5599</td>
<td>(\theta_{CC} = 0.254)</td>
</tr>
<tr>
<td>cosine</td>
<td>0.4200</td>
<td>0.0322</td>
<td>0.5777</td>
<td>(\theta_{cosine} = 0.200)</td>
</tr>
<tr>
<td>and-combination</td>
<td>0.5600</td>
<td>0.0136</td>
<td>0.6268</td>
<td>(\theta_{CC} = 0.254)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(\theta_{cosine} = 0.200)</td>
</tr>
<tr>
<td>or-combination</td>
<td>0.2800</td>
<td>0.0471</td>
<td>0.5108</td>
<td>(\theta_{CC} = 0.254)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(\theta_{cosine} = 0.200)</td>
</tr>
</tbody>
</table>

Test Results. The test results are presented in Table 9. As in training, the miss rates are
much higher than false alarms. As expected, the or-combination method is effective in
lowering miss rates. 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 standalone use of cosine and CC, respectively. Or-
combination $C_{Det}$ values are significantly smaller than those of and-combination ($p=0.0277$).
However, there is no statistical evidence supporting the difference between or-combination
and cosine; the same is also true for or-combination and CC. During training, the cost of CC
(0.5599) is smaller than cosine’s cost (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 training 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 the others: it decreases the
miss rate (or-combination’s miss rate is 33% lower than those of CC and cosine, 50% lower
than and-combination’s). As a result of decrease in miss rates, or-combination also yields the
lowest cost in terms of $C_{Det}$ value.

<table>
<thead>
<tr>
<th>Method</th>
<th>Miss Rate</th>
<th>False Alarm</th>
<th>$C_{Det}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.5313</td>
<td>0.0214</td>
<td>0.6359</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.3750</td>
<td>0.0244</td>
<td><strong>0.4947</strong></td>
</tr>
<tr>
<td>and-combination</td>
<td>0.6250</td>
<td>0.0103</td>
<td>0.6757</td>
</tr>
<tr>
<td>or-combination</td>
<td>0.2813</td>
<td>0.0355</td>
<td><strong>0.4550</strong></td>
</tr>
</tbody>
</table>

**TOPIC TRACKING: EXPERIMENTAL RESULTS**

To start TT some sample stories are needed about the topic to be tracked. In the
literature, usually between 1 and 4 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 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 is equal to
50% of average number of unique non-stopwords 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 is equal to 83% of average number of unique
use a similar approach for adaptive tracking.

**Comparative Evaluation of TT Methods**

**Training Results.** Figure 8 gives the static topic tracking DET plots for CC and cosine. A
comparison of Figure 8 and 7 shows that TT is considerably more effective than NED. 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 ($\theta_{adapt}$) is determined after obtaining the $\theta_{e}$ thresholds for CC and cosine. In static
TT 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 $p<0.05$). Or-
combination Min. $C_{Det}$ value is significantly smaller than that of CC ($p=0.092$); however,
there is no significant difference between or-combination and cosine.
<table>
<thead>
<tr>
<th>TT Sub Method</th>
<th>TT Method</th>
<th>Miss Rate</th>
<th>False Alarm</th>
<th>Min. C_{Det}</th>
<th>TT Threshold (θ_\text{on})</th>
<th>Adaptive Threshold (θ_\text{adapts})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>CC</td>
<td>0.0580</td>
<td>0.0130</td>
<td>0.1224</td>
<td>θ_{CC} = 0.097</td>
<td>N/A*</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>0.0462</td>
<td>0.0085</td>
<td>0.0881</td>
<td>θ_{cosine} = 0.099</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>And-combination</td>
<td>0.0425</td>
<td>0.0083</td>
<td>0.0824</td>
<td>θ_{CC} = 0.045</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>θ_{cosine} = 0.094</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or-combination</td>
<td>0.0462</td>
<td>0.0085</td>
<td>0.0881</td>
<td>θ_{CC} = 0.335</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>θ_{cosine} = 0.099</td>
<td></td>
</tr>
<tr>
<td>Adaptive</td>
<td>CC</td>
<td>0.0540</td>
<td>0.008</td>
<td>0.0910</td>
<td>θ_{CC} = 0.085</td>
<td>θ_{CC} = 0.400</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>0.0320</td>
<td>0.007</td>
<td>0.0650</td>
<td>θ_{cosine} = 0.076</td>
<td>θ_{cosine} = 0.300</td>
</tr>
<tr>
<td></td>
<td>And-combination</td>
<td>0.0260</td>
<td>0.007</td>
<td>0.0580</td>
<td>θ_{CC} = 0.027</td>
<td>θ_{CC} = 0.400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>θ_{cosine} = 0.069</td>
<td>θ_{cosine} = 0.300</td>
</tr>
<tr>
<td></td>
<td>Or-combination</td>
<td>0.0330</td>
<td>0.006</td>
<td>0.0650</td>
<td>θ_{CC} = 0.149</td>
<td>θ_{CC} = 0.400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>θ_{cosine} = 0.080</td>
<td>θ_{cosine} = 0.300</td>
</tr>
</tbody>
</table>

* N/A: Not applicable.

FIG. 8. Static TT training performance with CC and cosine similarity measures.

The adaptive topic tracking Min. C_{Det} values are better (lower) than those of static TT. Similar to the static case, and-combination provides the best effectiveness. Its Min. C_{Det} value is statistically significantly smaller than those of CC and cosine (p < 0.05 for both cases). It is close to be statistically significantly smaller than that of 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, 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 and-combination significantly improves the effectiveness with respect to CC, cosine, and or-combination methods (p < 0.05 for all cases). As in the training case, cosine provides an effectiveness (0.0842) statistically significantly smaller (p= 0.077) than that of CC (0.1277).

<table>
<thead>
<tr>
<th>TT Method</th>
<th>TT Sub Method</th>
<th>Static TT</th>
<th>Adaptive TT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Miss Rate</td>
<td>False Alarm</td>
<td>C_{Det}</td>
</tr>
<tr>
<td>CC</td>
<td>0.0355</td>
<td>0.0188</td>
<td>0.1277</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.0494</td>
<td>0.0071</td>
<td>0.0842</td>
</tr>
<tr>
<td>And-combination</td>
<td>0.0445</td>
<td>0.0071</td>
<td>0.0791</td>
</tr>
<tr>
<td>Or-combination</td>
<td>0.0494</td>
<td>0.0071</td>
<td>0.0843</td>
</tr>
</tbody>
</table>

In adaptive TT, or-combination provides the lowest cost (highest effectiveness). The cost of or-combination is 10% smaller than that of the and-combination (0.0461 vs. 0.0513); however, it is not statistically significantly smaller (p= 0.1121). The percentage cost reduction provided by 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 “and-combination” (0.0513) is close to be statistically significantly smaller (p= 0.0975) than that of CC (0.0699).

Note that in adaptive TT, in terms of best combination method we observe a contradiction between training and test results: during training, and-combination outperforms or-combination; during testing, the reverse is true. This can be explained by the fact that what we observe during training is a not “strongly” statistically significant (p= 0.0902, i.e. p value is relatively large and somewhat strongly significant –i.e., we have suggestive evidence). As indicated above, we have a similar observation for the comparison of 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 experiment with NED. While performing these experiments we use a) word stopping and the longest stopword list, 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) use Okapi; Franz, McCarley, Ward, & Zhu (2001) use a symmetrized version of the Okapi formula; and Brants, Chen, & Farahat (2003) use the Hellinger measure. The definitions of the Dice, Jaccard, overlap similarity measures can be seen in (Anderberg, 1973; 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 reports another version of Okapi that uses chronological term ranking (Troy & Zhang, 2007), which is outside the scope of this study.) We have already reported the CC and cosine results in the previous sections but they are 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%, provides the second best performance after cosine. The Jaccard measure introduces a cost increase of 4.2%. The others introduce a cost increase between 10.4 and 55.6%. In five out of seven cases (namely cosine, Dice, Hellinger, Jaccard, and Okapi measures), the best performance is observed with the lemmatizer-based stemmer LM. However, in two cases (CC and overlap similarity measures) the simple term truncation approaches (F5, F6) outperform LM. In four cases out of seven, it is better to use all terms of news stories for indexing. The use of cosine yields statistically significantly lower $C_{Det}$ values than two of the other similarity measures, namely CC and overlap (p < 0.05 for both cases).

We also experimented with all possible paired and- and or-combinations of these similarity measures. In the case of and-combination, the highest cost reduction is observed with the Okapi-overlap combination. It lowers the cost about -5% ($C_{Det}$ value of 0.5012 versus the best of the Okapi and overlap $C_{Det}$ values: 0.5249) when they are used alone. The other cases of the and-combination provide either small improvements or decreases in the effectiveness. In the case of 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 versus the best of the CC and cosine $C_{Det}$ value when they are used alone: 0.4948). 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 provide 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.

<table>
<thead>
<tr>
<th>Similarity Measure</th>
<th>Stemmer</th>
<th>Vector Size</th>
<th>Training Min. $C_{det}$</th>
<th>Test $C_{Det}$ (relative +% change)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>F6</td>
<td>60</td>
<td>0.5599</td>
<td>0.6359 (28.5)</td>
</tr>
<tr>
<td>Cosine</td>
<td>LM</td>
<td>All Terms</td>
<td>0.5777</td>
<td>0.4948 (0.0)</td>
</tr>
<tr>
<td>Dice</td>
<td>LM</td>
<td>All Terms</td>
<td>0.5669</td>
<td>0.5165 (10.4)</td>
</tr>
<tr>
<td>Hellinger</td>
<td>LM</td>
<td>All Terms</td>
<td>0.6207</td>
<td>0.5638 (14.0)</td>
</tr>
<tr>
<td>Jaccard</td>
<td>LM</td>
<td>All Terms</td>
<td>0.5664</td>
<td>0.5154 (4.2)</td>
</tr>
<tr>
<td>Okapi</td>
<td>LM</td>
<td>50</td>
<td>0.5424</td>
<td>0.5249 (2.1)</td>
</tr>
<tr>
<td>Overlap</td>
<td>F5</td>
<td>30</td>
<td>0.6573</td>
<td>0.7700 (55.6)</td>
</tr>
</tbody>
</table>

* Relative percentage change with respect to $C_{Det}$ value of Cosine.
However, none of the combination results is statistically significantly smaller than those of the standalone 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, & Wu (2007, Table 4) provide some state-of-the-art observations for the TDT3 corpora that includes documents in English and Mandarin. Their NED normalized $C_{Det}$ values for five different approaches are between “0.5413 and 0.6493” (average value is 0.5973). Our test phase $C_{Det}$ values (Table 12) for the standalone use of the seven similarity measures are between “0.4948 and 0.7700” (average value is 0.5745). Allan et al. (2003, Figure 3) also provide evidence in the same direction 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 is generated by using the web site of “Türk Dil Kurumu” (TDK) – Turkish Language Association-. The TDK website (http://www.tdk.gov.tr) provides a dictionary of person names. The personnel, student, and high school student information databases of Bilkent University are also used for human names. For location names, address records of personnel and student information databases of Bilkent University are used, city (in Turkish “şehir”), county (ilçe) and district (semt) names are also inserted into the database. Organization names table is manually created by using frequently used organization names such as TRT, TÜBİTAK, MEB, etc. 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 & Allan (2004) also have 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 have already 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 if used it has a degenerative effect (Kardas, 2009, pp. 57-58). However, if 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 start to get a performance not as good as window size of 14 days but close to it. Finally, using a considerably larger window size of 100 days the effectiveness level reaches 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 worsen again. For the cosine measure, we have 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 after that). 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 are also comparable to those in other languages; however, for NED there is still a large room for improvement (Allan et al., 2000; Allan et al., 2003).

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 miss rates 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 stopword list: the use of stopword list 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.

Acknowledgements

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 paper. 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 the grant number 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 and therefore no official endorsement should be inferred.

References


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


