# Novelty Detection for Topic Tracking

**Cem Aksoy[1], Fazli Can, and Seyit Kocberber**
*Bilkent Information Retrieval Group, Computer Engineering Department, Bilkent University, Ankara, Turkey 06800. E-mail: canf@muohio.edu; ca64@njit.edu; canf@cs.bilkent.edu.tr; seyit@bilkent.edu.tr*

**Multisource web news portals provide various advantages such as richness in news content and an opportunity to follow developments from different perspectives. However, in such environments, news variety and quantity can have an overwhelming effect. New-event detection and topic-tracking studies address this problem. They examine news streams and organize stories according to their events; however, several tracking stories of an event/topic may contain no new information (i.e., no novelty). We study the novelty detection (ND) problem on the tracking news of a particular topic. For this purpose, we build a Turkish ND test collection called *BilNov-2005* and propose the usage of three ND methods: a cosine-similarity (CS)-based method, a language-model (LM)-based method, and a cover-coefficient (CC)-based method. For the LM-based ND method, we show that a simpler smoothing approach, Dirichlet smoothing, can have similar performance to a more complex smoothing approach, Shrinkage smoothing. We introduce a baseline that shows the performance of a system with random novelty decisions. In addition, a category-based threshold learning method is used for the first time in ND literature. The experimental results show that the LM-based ND method significantly outperforms the CS- and CC-based methods, and category-based threshold learning achieves promising results when compared to general threshold learning.**

## Introduction

The Internet has changed the news industry (*The Economist*, 2011). Most newspapers and news agencies provide news on their web pages. News portals work as a news aggregator and gather, merge, and organize news articles obtained from various sources. Multisource news portals provide various advantages such as richness in news content and an opportunity to follow event developments from different perspectives. In addition, it is practical to follow different news sources from a single web page. Google News (http://news.google.com) is a well-known commercial news portal example. It offers many services such as information retrieval, personalized information filtering, and news clustering. Research-oriented examples include NewsBlaster (McKeown et al., 2002) and NewsInEssence (Radev, Otterbacher, Winkel, & Blair-Goldensohn, 2005), each of which provides clustering and summarization services over the news.

As the number of sources and events increase, news readers may be overloaded with information and thus may face difficulty in finding news related to their interests. Different organizational techniques have been employed for more effective, efficient, and enjoyable browsing. Studies on new-event detection and topic tracking aim to organize news with respect to events or topics. In topic detection and tracking (TDT), an event is defined as a happening that occurs at a given "place and time, along with all the necessary preconditions and unavoidable consequences" (Topic Detection and Tracking Initiative, 2004, p. 4). For example, the Fukushima Daiichi nuclear accident of March 11, 2011 is an event starting a new topic. In TDT studies, a topic is defined as "a seminal event or activity with all directly related events and activities" (Topic Detection and Tracking Initiative, 2004, p. 4). So, a topic can be about the developments related to a specific nuclear accident, and not all or other nuclear accidents (e.g., Idaho Falls and Chernobyl are different topics).

Various problems were attacked by the Topic Detection and Tracking research initiative (Allan, Carbonell, Doddington, & Yamron, 1998). One of these, 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 one to four) known to be on the same topic.

This study follows our earlier studies on information retrieval on Turkish texts (Can, Kocberber, Balcik, et al., 2008) and new event detection and TT in Turkish (Can et al., 2010). An overview of Turkish, the language mainly used in the republic of Turkey, is provided in the first study and is not repeated here. The second study shows that it is possible to reach a TT success rate which is high enough to use in

Received May 20, 2011; revised September 12, 2011; accepted September 28, 2011

[1]Present address: Computer Science Department, New Jersey Institute of Technology, University Heights Newark, NJ 07102.

 DOI: 10.1002/asi.21697

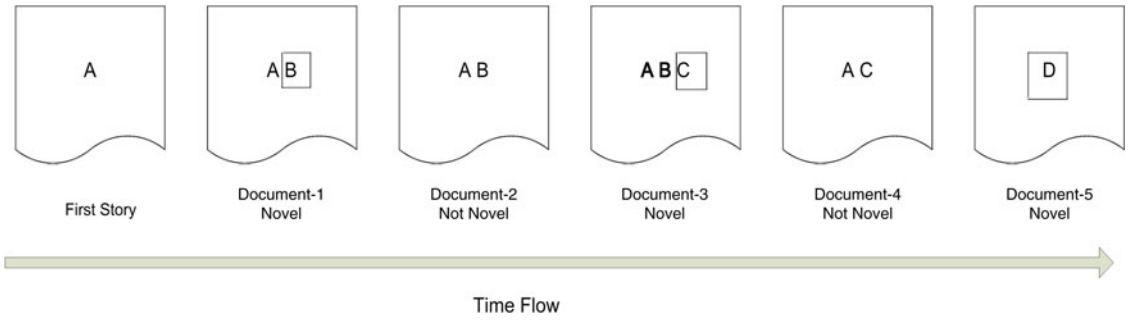


FIG. 1. Illustration of ND in context of topic tracking.

operational news web portal environments (Can, Kocberber, Baglioglu, et al., 2008; Öcalan, 2009). However, in real-life applications, TT by itself may not be sufficient since many tracked news streams of a topic contain no novel (i.e., new) information with respect to earlier ones. In such environments, documents with novel information can be detected and made more noticeable using a timeline. For example, Allan, Aslam, et al. (2003) showed novelty detection (ND) as a necessary complement to real-world filtering systems.

ND may be defined as finding data which contain novel characteristics with respect to some other, mostly earlier, data. It has been studied in many domains at different scales with slightly differing problem definitions. In signal processing, the task is to identify new or unknown data which has not been encountered during the training process (Markou & Singh, 2003). This task is also named as outlier detection (Hodge & Austin, 2004). In text processing, ND has been studied in different scales with different definitions: event-based or information-based. The purpose of event-based ND is to find novelty at the event scale. This also can be explained as detecting the initial reporting of a new event. Information-based ND tries to find pieces of text which contain some information which was not contained in some reference text (discussed later). In this work, we use the novelty definition used in information-based ND studies. Given the tracking news of a topic, we try to identify documents containing novel information not covered in any of the previous documents. (In the article, the words "news," "story," and "document" as well as "effectiveness" and "performance" are used interchangeably.) Novelty decision is given for documents; however, this decision can be made by analyzing the document sentences. In Figure 1, an illustration of the ND problem in this context is given. Let A, B, C, and D represent different information contained by the documents. Rectangles show the piece of information which causes the document to be regarded as novel. The first story is novel by default. Document 1 is novel because it reports information not reported earlier (Information-B). Document 2 is not novel because it contains no novel information: Both A and B were reported earlier. Document 3 reports Information-C and is novel. Document 4 is not novel, and Document 5 is novel. Document 4 shows another important characteristic of the ND problem: It is different from near-duplicate detection (Chowdhury, Frieder, Grossman, & McCabe, 2002; Varol, Can, Aykanat, & Kaya, 2011). Although both ND and near-duplicate detection aim to eliminate redundancy, Document 4 is neither a near-duplicate of any of the previous documents nor is it novel. This shows that ND should be handled in a different manner than near-duplicate elimination.

Relevancy and novelty are contradictory in some sense that sentences/documents should be similar to previous ones in order to be relevant but they need to be dissimilar in order to be novel. Since these two tasks are conflicting, they should be evaluated separately (Zhang, Callan, & Minka, 2002). In this work, we will track documents of a topic (Aksoy, 2010), so all of the documents are assumed to be relevant to the topic. Even though we work on TT, the methods studied in this article can be applied in other application domains that involve streaming data, such as information filtering, financial analysis, intelligence applications, patient watch, and so on.

### *Contributions*

In this article, we

- Give the details about the construction and characteristics of a large ND test collection, *BilNov-2005* (Bilkent ND test collection). It contains 59 annotated events. *BilNov-2005* (2010) is available to other researchers as the first test collection prepared for ND studies for TT in Turkish.
- Propose the usage of three different ND methods on TT and similar applications: a cosine-similarity (CS)-based ND method, a language-model (LM)-based ND method, and a cover-coefficient (CC)-based ND method. We show that the LM-based ND method significantly outperforms the other two methods statistically, is highly successful, and can be used in real-life applications.
- Introduce a baseline for ND studies that quantifies the performance of an ND system with random decisions.
- Show that when compared with a general threshold learning approach, our category-based threshold learning approach yields promising results even with small amounts of information for the categories.
- Demonstrate that our results are comparable with those in English based on sentence-level ND experiments [using the Text Retrieval Conference (TREC) 2004 novelty track test collections] (Soboroff, 2004).

The rest of the article is organized as follows. First, we review ND studies by categorizing them as event-based,


DOI: 10.1002/asi

information-based, and other applications. Next, we explain construction details of the ND test collection *BilNov-2005* and the ND methods. We then present evaluation measures for ND and the effectiveness assessments of the ND methods investigated in this study. Finally, we conclude with a summary of our findings and some future research avenues.

## Related Work

Li and Croft (2008) categorized ND studies into three classes: event level, sentence level, and other applications. We follow a similar approach by naming the categories as event-based, information-based, and other applications.

### *Event-Based ND*

The new-event-detection problem is mainly introduced in the Topic Detection and Tracking Initiative research initiative (Allan et al., 1998). Different techniques are utilized to attack the event detection; that is, the first story detection (FSD) problem. Clustering is widely used to cluster news articles which report the same event into the same cluster. An incoming-story's similarities to the previous clusters are calculated, and if the story is dissimilar to all of the previous clusters to an extent, it starts a new cluster and is labeled as a new event (Manning, Raghavan, & Schütze, 2008, p. 362). This is similar to the single-pass clustering explained in van Rijsbergen (1979, p. 52). In this approach, efficiency degradation may occur as the number of clusters increase. Yang, Pierce, and Carbonell (1998) proposed a sliding time window concept in which an incoming story is only compared to the members of a time period, thereby decreasing the number of comparisons. They also utilized a time-decay function to lessen the influence of older documents.

The use of named entities in TDT systems also was examined. Yang, Zhang, Carbonell, and Jin (2002) introduced a two-level scheme in which they first classify incoming stories to broader topics such as "airplane accidents," "bombings," and so on before performing new-event detection. After this classification, stories are compared to the local history of the broader topic instead of all documents processed by the system. This increases the efficiency with respect to normal FSD systems, which compare incoming stories with all of the previous documents. In addition, named entities are given weights specific to the topics. This is one of the rare studies in which employing named entities significantly increases performance, which may be due to the two-level scheme. Kumaran and Allan (2004) and Can et al. (2010) reported no significant improvement when named entities are used, and stated that this may be caused by the test collections used not being conducive to the usage of named entities.

Event detection also is addressed in Automatic Content Extraction (ACE) workshops organized by National Institute of Standards and Technology (NIST) (ACE, 2005).

### *Information-Based ND*

Information retrieval systems rank the documents in a collection in terms of relevance to a query and provide the ranked list to the user. As the number of documents increases, redundant information increases as well. To handle such collections with redundant information, a search system that detects relevancy and novelty is required.

The NIST organized TREC novelty track workshops between 2002 and 2004 (Harman, 2002; Soboroff, 2004; Soboroff & Harman, 2003). In these workshops, two problems were defined for a list of documents (split into sentences) that are relevant to a query. These are:

- *Relevant Sentence Retrieval:* This problem aims to find sentences relevant to the query. Sentence retrieval is considered to be different from document retrieval because sentences are shorter than documents (Soboroff & Harman, 2005). Since they contain less text, systems that work on sentences may be less reliable. Despite this potential problem, taking sentences as the unit of retrieval enables adjusting sentence-level decisions to different levels of texts.
- *Novel Sentence Retrieval:* This problem aims to identify relevant sentences which contain new information with respect to the previous relevant sentences both in the same document and in the previous documents. This definition constrains novel-sentence-detection algorithms to run in an incremental way in which every sentence adds some knowledge which should be examined to decide the novelty of the next sentence. Another important point of novel-sentence detection is that it should be done over relevant sentences because new information in irrelevant sentences should not be presented to the users.

The test collections used in TREC novelty tracks comprise about 50 topics, each containing a query and 25 relevant documents. In TREC 2004, some irrelevant documents are included in the topics to make the task more challenging. In the Novelty 2002 track, the documents are given in the order of relevance; in 2003 and 2004, the documents are processed in chronological order, which is more appropriate for the nature of ND. Documents were split into sentences by the NIST, and the annotators select the set of relevant sentences, and within the set of relevant sentences, then they select the novel sentences (Soboroff & Harman, 2005). Performance evaluations are conducted over these ground truth data. F-measure is used for assessment (van Rijsbergen, 1979).

There were four different tasks with varying quantities of training data:

- *Task 1:* Given the set of all documents and the query, find all relevant and novel sentences.
- *Task 2:* Given the set of relevant sentences, find all novel sentences.
- *Task 3:* Given the relevant and novel sentences for the first five documents, find relevant and novel sentences in the remaining 20 documents.
- *Task 4:* Given all relevant sentences and novel sentences for the first five documents, find novel sentences in the remaining 20 documents.

In the following, we consider only related work on novel-sentence-retrieval methods since relevance detection is beyond the scope of this work.


DOI: 10.1002/asi

In TREC novelty tracks, a very simple but intuitive method, *New Word Count*, is one of the most successful methods (Larkey, Allan, Connell, Bolivar, & Wade, 2002). In this method, the novelty of sentences is based on the number of new words that they contain. A "new word" in this context is a word that is encountered for the first time. This method needs a threshold value for making a novelty decision.

Similarity measures also are utilized for ND. Basically, a sentence is compared to all previous sentences, and if the similarities to all of the previous sentences are below a threshold, the sentence is labeled as novel. This idea is adapted from FSD in TDT (Papka, 1999). M.-F. Tsai, Hsu, and Chen (2004) used the CS measure for similarity calculation. Instead of comparing a current sentence with all previous sentences one by one, Eichmann et al. (2004) compared it with a knowledge repository consisting of all previous sentences. Zhang et al. (2002) claimed that since novelty is an asymmetric property, symmetric similarity/distance measures may perform poorly in ND. In their study however, CS, which is a symmetric measure, was successfully utilized. Cheng (2005) also uses CS as a novelty measure for applying ND on TT. To the best of our knowledge, Cheng's work is the only application of ND on TT so far.

LMs also are employed for novel-sentence detection. Kullback–Leibler (KL) divergence is a measure that calculates the difference between two probabilistic distributions. It can be used for measuring the dissimilarity of two LMs (Zhang et al., 2002). Two different approaches are followed during calculation of KL divergence (Allan, Wade, & Bolivar, 2003): an aggregate and a nonaggregate approach. In the aggregate approach, for a sentence, KL divergence of the sentence LM and an LM constructed from all of the previously presumed relevant sentences is calculated. The novelty score of a sentence is proportional to this KL-divergence value. In the nonaggregate approach, separate LMs are constructed for each sentence, and the novelty of a sentence is found as the minimum KL-divergence value calculated between the sentence LM and all of the previously presumed relevant sentence LMs.

Different smoothing approaches are used for LM, such as Jelinek–Mercer and Dirichlet smoothing (Zhai & Lafferty, 2004), to overcome the problem of having terms with zero probabilities. In addition to these, a mixture model was proposed by Zhang et al. (2002). It tries to model every sentence as a set of words generated by three different models: a general English model, a topic model, and a sentence model.

Li and Croft (2008) addressed the ND problem within the context of question answering. They defined novelty as new answers to a possible information request made by the user's query. Queries are converted into information requests. Named-entity patterns such as person ("who") and date ("when") are used as question patterns. Then, sentences that have answers to these questions are extracted as novel ones. Problems arise in opinion topics, whose queries do not include such patterns. Different patterns, such as "states that," are proposed for opinion topics. In addition, a detailed information-pattern analysis of sentences in TREC novelty data was given in the article.

### *Other Applications*

ND techniques may be applied in many areas such as intelligence applications, summarization, and tracking of developments in blogs and patient reports.

Zhang et al. (2002) extended an adaptive filtering system for redundancy elimination. Documents to be delivered for a filtering profile are processed by a redundancy-elimination tool. Documents that are redundant (given the previously delivered documents) are eliminated. Experiments on different measures were conducted in their study. The best performing methods were a CS-based method adapted from FSD and another based on the mixture of LMs.

ND at the sentence level has many similarities with that of summarization studies. In both of them, only the necessary sentences should be delivered to the user (Sweeney, Crestani, & Losada, 2008). In summarization, there also is a necessity to compress the given text, which is not valid for ND studies in TREC. This may be explained as follows: If a newer sentence contains the information provided in a previous sentence, but also provides some new information, both of the sentences are labeled as *novel* in ND. However, because of compression concerns, only the latter sentence may be contained in the summary. A subtopic of the summarization area, temporal summarization, aims to generate a summary of a news stream, considering the previous summaries and providing an update to the previously delivered summary. Allan, Gupta, and Khandelwal (2001) defined the usefulness (similar to relevancy) and novelty of sentences, and tried to extract novel and useful sentences. Language modeling was used with a very simple smoothing approach. In addition, update summarization is a similar problem which is piloted in Document Understanding Conference 2007 and continued in Text Analysis Conferences 2008 and 2009 (Dang & Owczarzak, 2008; Text Analysis Conference, 2009). The aim of update summarization is to generate a summary for a set of documents under the assumption that another set of documents already has been read by the user.

Temporal text mining deals with analyzing temporal patterns in text. In Mei and Zhai (2005), evolutionary theme patterns are discovered. As an example given in the paper, in a text stream related to the Asian tsunami disaster, the aimed themes are "immediate reports of the event," "statistics of death," "aid from the world," and so on. In addition, a theme-evolution graph is extracted in which transitions between themes are shown. LM also was utilized in their study. Parameters of the probabilistic models are estimated by expectation-maximization algorithm (Moon, 1996).

## ND Test Collection Construction and *BilNov-2005*

In this section, we report the construction details of the first Turkish ND test collection, *BilNov-2005* (Aksoy, 2010). To the best of our knowledge, it also is the first large-scale ND test collection constructed for "topic tracking" in any


DOI: 10.1002/asi

language; the first one by Cheng (2005) contains 16 events. *BilNov-2005* is based on the TDT test collection *BilCol-2005* (Can et al., 2010). Information on the annotated topics is given in Appendix A Table A1. In that table, the first row is for a topic about an accident that took place in Kars, a city in the eastern part of Turkey; this topic had 20 tracking stories. The dates of the first story and last story are May 28 and December 16, respectively. (All dates for all topics are from Year 2005.) The news categories are the same as defined for the Topic Detection and Tracking Initiative (2004) studies. The Topic Detection and Tracking Initiative defined 13 categories, and in *BilNov-2005*, some of these categories contain no news topics in that category, such as the category elections. On the other hand, for example, the category "scandals/hearings" contains six topics.

*Selection of Topics Used in BilNov-2005*

The *BilCol-2005* TDT test collection, the base of *BilNov-2005*, consists of 80 topics with an average of 73 tracking news identified in a news stream that contains 209,305 stories after eliminating duplicate and near-duplicate documents (Can et al., 2010). Although the average number of tracking stories is 73, it contains topics with only a few tracking stories (as low as 5) and topics with many (as high as 454) tracking stories. Our experience has shown that topics with a large number of tracking stories are difficult to annotate for novelty since with each additional document ND annotation time, the extent of information that should be remembered increases. On the other hand, topics with a very small number of tracking stories are not appropriate for assessing ND methods; such topics are not challenging enough to use in performance evaluation because they do not involve many decisions to make. Accordingly, 59 topics from *BilCol-2005* containing at least 15 tracking documents were chosen, and for topics with 80 or more tracking stories, their first 80 documents were used.

*Annotation Process*

Documents were examined by human annotators/assessors in time sequence. (Each document has a timestamp.) The annotators, all native speakers of Turkish, are mostly graduate students of computer engineering and a few colleagues. We worked with 38 different annotators. The annotators have a different number of topics assigned to them, but we tried to make a balanced assignment to each annotator in terms of the total number of documents to be assessed. The annotations are carried out by using a web interface, and the annotators are asked to use their judgment about the novelty of information provided in news articles.

An annotator reads the first story of a topic and then reads the tracking documents in time order. After reading a tracking document, the annotator decides whether it is novel (i.e., contains new information) compared to all earlier documents of the same topic. Annotators are allowed to reexamine any annotated document and change their decision. They also are allowed to take breaks. At the end of the annotation process, they enter the amount of time they spend during annotation without including the breaks (if any). The annotation times span between 15 and 163 min, with an average, median, and standard deviation of 61, 53, and 35 min, respectively. The novelty decision time needed for each document in terms of average, median, and standard deviation was 1.21, 1.13, and 0.36 min, respectively.

In similar applications, generally multiple annotators are used for the assessment of the same item. These multiple judgments may be used separately to observe different points of views; however, in general, a single ground truth data is obtained by combining them. In our study, each topic was assessed by two annotators. For combining judgments, a majority voting approach would not work with two decisions. Furthermore, such an approach removes the opinions of different annotators. In some studies, in cases of disagreement, annotators had been asked to work together to decide on one of the decisions. In ND, this reevaluation process is rather difficult since it may, and in most cases does, require the reexamination of all documents from the very first story because the reason why a document is tagged as novel or not novel can be forgotten after a certain amount of time. The difficulty also comes from the fact that the annotation process is quite boring (for a discussion of similar kinds of difficulties in a similar novelty test set creation in information filtering to Zhang et al., 2002). In such tasks, reevaluations can make the annotations even less reliable since some decisions may unconsciously become almost arbitrary to end the annotation process.

*Combining annotations: Optimistic and pessimistic ground truths.* We follow a similar approach to that of Zhang et al. (2002) by combining the decisions of the annotators. In their work, Zhang et al. (2002) instructed the annotators to give novelty decisions at three levels: "absolutely novel," "somewhat novel," and "not novel." Later, they conducted experiments with these data by taking "somewhat novel" ones as "novel" in one configuration and as "not novel" in the other configuration. This setup enables them to evaluate their systems in terms of sensitivity to strictness of novelty decision. If we neglect possible annotator mistakes, the disagreement between the decisions is probably caused by different interpretations of novelty (discussed later). So, if we combine decisions of annotators in two different ways, we would be able interpret novelty in different dimensions. These two configurations are defined as follows.

- *Optimistic ground truth:* When two annotators are in disagreement, we choose the decision which is more optimistic about novelty of the document. In other words, if one of the decisions is "novel," the optimistic ground truth label also is novel. This is similar to logic function, *OR*, if we consider novelty as 1, if any of the decisions is 1, the optimistic ground truth also is 1.
- *Pessimistic ground truth:* In this ground truth data, contrary to optimistic ground truth, ground truth label is novel if and only if both of the annotator judgments are novel. This is similar to logic function, *AND*, causing the ground truth label to be 0 if one of the decisions is 0 (i.e., not novel).


DOI: 10.1002/asi

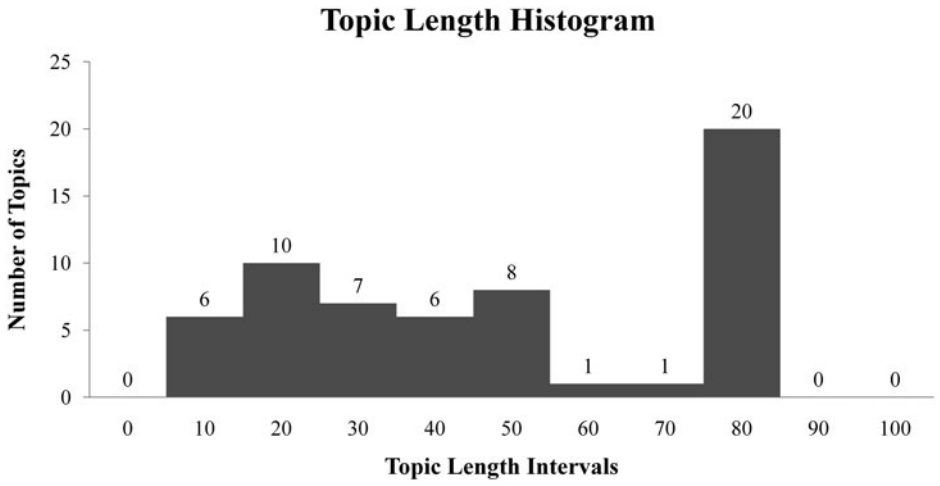


FIG. 2. Histogram illustrating the distribution of topic lengths in *BilNov-2005*.

*Quality Assessment of Annotations*

Construction of experimental test collections in information retrieval and related studies requires dealing with lots of data and several assessments. It is difficult to examine these one by one to evaluate their correctness or appropriateness for the task for which the collection was built. During or after annotations, some quality-control techniques generally are applied to both data and judgments (Conrad & Schriber, 2006). With the help of these techniques, errors about a test collection may be corrected. In our case, inappropriate topics and topics with unreliable annotations may be identified and reassessed.

In annotations, we like to have a "considerable amount of agreement" among the assessors of a given topic. In other words, we understand that assessors may have "some disagreement" in their decisions. In ND, among other things, disagreements among annotators come especially from the nature of the concept of novelty: Sometimes it is very concrete, and sometimes it can be quite subjective and opinion-based.[2] This flexibility gives an opportunity of representing different human opinions (for a similar approach, see Soboroff, 2004). On the other hand, we do not want to accept two ND assessments regarding a certain topic that involve disagreements at the level of arbitrariness or randomness. Therefore, during the construction of *BilNov-2005*, for some topics the annotations were thrown away and were repeated from the very beginning by two completely different assessors.

In the following section, we present the details about the quality analysis that we performed in terms of topic lengths, novelty ratios, and interannotator agreement.

*Analysis of topic lengths.* Topic lengths are important for an ND test collection. A test collection built from short topics (i.e., events that involve a small number of tracking documents) may not result in a reliable assessment environment since such topics can be limited in terms of the number of observations, case variety, and test conditions that they provide. In addition, choosing topics of the same length has the potential of hiding some possible biases of ND methods. Figure 2 shows that *BilNov-2005* consists of topics with a variety of lengths and therefore provides a rich test environment.

*Analysis of novelty ratios.* Novelty ratio of documents for a particular topic is defined as the ratio of the number of documents labeled as *novel* to the total number of tracking stories for the topic. It is desirable to use a test collection with a wide variety of cases in terms of novelty ratios to have a variety in test collections (F.S. Tsai, Tang, & Chan, 2010). We depict the distribution of novelty ratios for both ground truth data in Figure 3, the novelty detail of the individual topics is given in Table A1. Figure 3 shows that *BilNov-2005* topics have a wide variety in terms of topic-novelty ratios.

*Interannotator agreement.* Reliability of a ground truth data constructed from the decisions of different annotators depends on the agreement between annotators. Kappa coefficient is widely used for measuring interannotator agreement (Cohen, 1960). Its value ranges between −1.00 and 1.00, and its formula is given in the following equation.

$$\kappa = \frac{Agr - E(Agr)}{1 - E(Agr)}.$$

[2]The subjectivity of novelty shows itself especially in the novelty interpretation of human annotators for small details. For example, while reporting an accident, a document may give the place of an accident in terms of the city in which it takes place whereas another document also may provide the neighborhood information. The novelty, or perhaps more correctly, the "significance" of this information may have different value for different people. Another example can be given in terms of quantitative information. For example, consider a news article "Sidney Lumet dies (1924, 2011) …" and consider a tracking article which reads "Sidney Lumet dies. He was 86…." For people who are not good at numbers, the age information may be interpreted as new information. Moreover, novelty assessment of long stories can be inevitably error-prone, especially if they contain small details: Due to the overwhelming effect of too many words, it becomes easier to miss or misinterpret details. In some other cases, a news article reporting known facts with different words or summarizing the course of event development can be erroneously interpreted as new.


DOI: 10.1002/asi

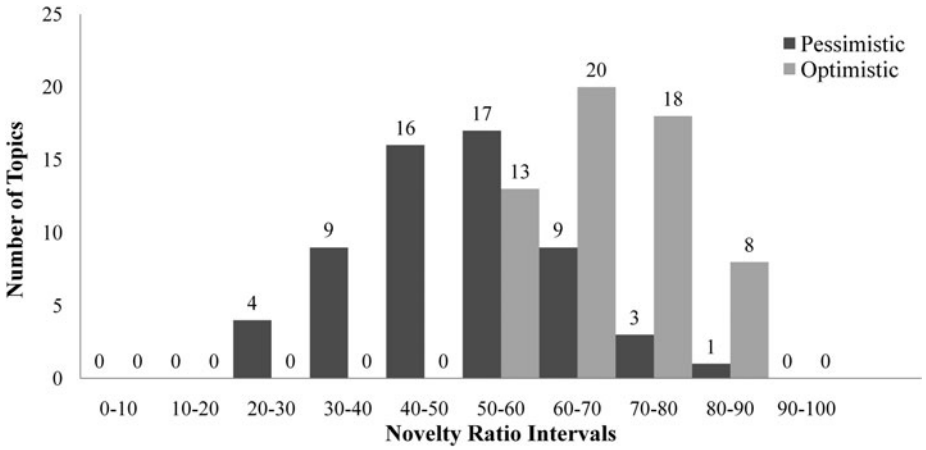


FIG. 3. Distribution of novelty ratios (in percentages) in *BilNov-2005*.

TABLE 1. Example case for kappa calculation between Annotators A and B.

| | B | | |
|---|---|---|---|
| Annotators' judgments | Novel | Not novel | Total |
| A | | | |
| Novel | 35 | 5 | 40 |
| Not novel | 40 | 20 | 60 |
| Total | 75 | 25 | 100 |

In this formula, *Agr* stands for the observed agreement between the annotators, *E*(*Agr*) is the expected agreement, which is calculated by using the individual probabilities of the annotators. In the denominator, *E*(*Agr*) is subtracted from 1 because 1 is the maximum value that an agreement can take, so this takes the role as a normalization factor (Jain & Dubes, 1988, p. 175). This way, we are correcting the statistics *Agr*. Kappa coefficient takes values less than or equal to 0.00 for cases where there is not an agreement more than the expected case, and its value is $-1.00$ when there is perfect disagreement below chance. In the case of perfect agreement, it takes the value 1.00.

An example case is given in Table 1. Rows represent the decisions of Annotator A and columns represent Annotator B. The expected agreement between the annotators is calculated as $0.75 \times 0.4 + 0.25 \times 0.60 = 0.45$. This is simply the sum of probabilities of cases where both annotators label the document as *novel* or *not novel*. The probabilities are obtained by their assessments. Agreement between A and B, *Agr*, is the sum of diagonal values which are the documents labeled as both *novel* or *not novel*. Therefore, $\kappa = [(0.35 + 0.20) - 0.45]/(1 - 0.45) \approxeq 0.18$.

In *BilNov-2005* judgments, the average kappa coefficient is 0.63. This value stands for a substantial agreement according to intervals given by Landis and Koch (1977). In addition, we performed the statistical test proposed by Conrad and Schriber (2006) which shows that the observed kappa value is significantly different from 0 with $p = 0.002$. It indicates that the agreements are significantly larger than the expected cases. In other words, agreement we observe in the annotations is not by chance.

## ND Methods

In this section, our proposed ND methods are explained. The CS- and LM-based approaches are adapted from ND literature (Allan, Wade, & Bolivar, 2003).

### *Category-Based Threshold Learning and Cross-Validation*

We utilize cross-validation for reporting our system performance since all of our methods have some parameters, and these should be optimized. In this study, motivated from Yang et al. (2002), we also try category-based threshold learning and compare the results of general threshold learning with category-based threshold learning. Yang et al. studied running FSD on a local history of documents based on a category, instead of all of the previous documents. Our motivation here is that each topic has a different category (e.g., sports news, accident news, etc.), and each of these categories possibly has a different novelty model. For example, intuitively, one would expect to see more rapid, but small, developments in an accident topic while in a topic related to politics, it may take days for the topic to become mature. Therefore, we hypothesize that while learning a threshold for a topic, if we use only topics from the same category in the training phase, system performance can be increased. In our test collection, there are 11 different categories (e.g., accidents, financial news, etc.) with two or more topics (see Table A1). We experiment with category-based threshold learning using these categories. For general threshold learning, we use 30-fold cross-validation, and for category-based threshold learning, we use leave-one-out cross-validation.

### *ND Methods*

*Baseline–random ND.* Systems which randomly give their decisions are widely used as a baseline in many problem


DOI: 10.1002/asi

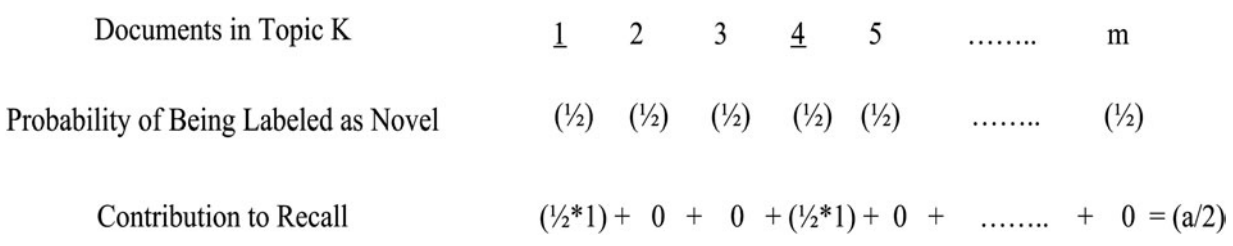


FIG. 4. Calculation of expected performance of random baseline.

areas (Jain & Dubes, 1988). In new TDT studies, it is traditional to compare the performance of a system with random performance (Fiscus & Doddington, 2002). A method's decisions are justified to be different from random decisions by comparing the system with a random baseline.

In ND context, the random baseline method gives novel/not-novel decisions with a probability of 0.5 without examining the contents of a document. To evaluate the random baseline, expected performance of such an approach should be found. This can be done by considering all novel/not-novel assignment configurations, calculating performance of the specific case, multiplying the performance of the case by the probability of occurrence of the case, and summing up this for all cases. We generalize this calculation with the help of the example given in Figure 4.

Let $K$ be a topic with $m$ documents, as in Figure 4, and $a$ be the number of novel documents in these $m$ documents. The first row of the figure shows the documents in which novel ones are underlined. The second row shows the probabilities of each document being labeled as *novel*. As stated earlier, this probability is 0.5 for all documents in random baseline. The third row shows the contribution of each document to recall if it would be in the set of documents returned by the system. Not-novel documents obviously do not make any contribution to both precision and recall. Novel documents will have one contribution to the measures; they can be involved in the set with 0.5 probability, so in the expected case, the sum will be $(a/2)$. Thus, we can derive recall as $R = (a/2)/a = \frac{1}{2}$. However, for precision, the contribution of a document is not only to the numerator part of the formula; the denominator part of a precision formula also increases (recall calculation can be done easily as we since the denominator part of recall is constant, $a$.) So, we derive a general formula for precision calculation for a topic with $m$ documents and $a$ novel documents where $a > 1$, which can be seen in the following equation. In the equation, the term $\binom{a}{i}\binom{m-a}{j}$ stands for the number of cases where $i$ novel documents can be chosen correctly from $a$ novel documents and where $j$ documents can be chosen from $(m - a)$ not-novel documents. Precision at this case is $\frac{i}{i+j}$, which is equal to the ratio of the number of novel documents in the set of returned documents to the total number of returned documents. The denominator $2^{\mathrm{m}}$ is the number of total cases. (It also might be taken as $2^{\mathrm{m}} - 1$ since in the case where no documents are returned, precision is not defined, but we neglect this.)

$$Precision = \frac{\sum_{i=1}^{a}\sum_{j=0}^{m-a}\binom{a}{i}\binom{m-a}{j}\frac{i}{i+j}}{2^m}$$

*CS-based ND.* In many text-based studies, the problem is usually reduced to accurately calculating the similarities between some pieces of texts and giving a decision based on these similarity values (generally with the help of a threshold value). CS is one of the most frequently used similarity measures in information retrieval. Its geometrical interpretation is the cosine of the angle between two vectors. The texts to be compared are initially converted into a vector-space model (Salton, 1989, pp. 313–326). In this model, every unique term is represented by a dimension in the vectors, and the values of these dimensions are obtained by a term-weighting function. *TF-IDF* function is very widely used as a term-weighting function in which *TF* indicates term frequency and *IDF* is the inverse document frequency. The function basically tries to give higher importance to the terms that occur frequently in a specific document, but not in all documents. In this study, we use raw *TF* values for term weighting because of unfavorable initial results obtained with the *TF-IDF* function. CS tends to give good results even with just raw term frequencies. Similar observations were reported in Allan, Lavrenko, and Swan (2002).

The following formula gives the CS calculation. In the numerator, the dot product of the vectors $w_i$ and $w_j$ is calculated by summing the multiplication of the corresponding dimensions. The denominator is a normalization factor which consists of multiplication of lengths of both of the vectors. $N$ is the number of dimensions in both vectors.

$$CosSim(d_1, d_2) = \frac{\sum_{k=1}^{N} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{N} w_{ik}^2 \cdot \sum_{k=1}^{N} w_{jk}^2}}.$$

Our CS-based method is adapted from FSD; document $d_t$ arriving at time $t$ is compared to all of the previous tracking documents, and if its CS to *any* of the previous documents is greater than the threshold value (obtained by training), the document is labeled as *not novel*; otherwise, the document is labeled as *novel*. In other words, a smaller threshold


DOI: 10.1002/asi

value implies that a smaller number of new documents will be classified as novel.

*LM-based ND.* Probabilistic models have been incorporated in information retrieval for over 4 decades (Zhai & Lafferty, 2004). These models try to estimate the probability that a document is relevant to the user query. Ponte and Croft (1998) introduced a simple probabilistic approach based on language modeling. This model, unlike its predecessors, does not have any prior assumptions on documents such as the case in a parametric model. Maximum likelihood estimate (MLE) of probability of term $t$ being generated from the distribution of document $d$ as introduced by Ponte and Croft is given in the following equation.

$$P_{MLE}(t|\theta_d) = \frac{tf(t,d)}{|d|}.$$

In the equation, $tf(t,d)$ is the term-frequency function, which gives the number of occurrences of $t$ in document $d$, and $|d|$ is the length of the document, which is the number of tokens in $d$. MLE formula basically gives probabilities to the terms which are proportional to their frequency in the document. If a term does not occur in the document, its probability is estimated as zero with MLE. This is a very strict decision and generally does not reflect the true probability of the term.

Smoothing approaches aim to correct the abnormalities of MLEs that assign zero probabilities to unseen terms. Especially when estimating a model with a limited amount of text, smoothing makes a significant contribution toward the model's accuracy (Zhai & Lafferty, 2004). Allan et al. (2001) applied smoothing in a simple way by adding 0.01 to the numerator of $P_{MLE}(t|\theta_d)$ and multiplying the denominator by 1.01. This approach helps to overcome problems caused by unseen terms; however, it does not offer a good estimate of the probability. In this study, we will experiment with two different smoothing approaches: Bayesian smoothing using Dirichlet priors and Shrinkage smoothing (Allan, Wade, & Bolivar, 2003; Zhai & Lafferty, 2004).

**LM-based ND: Bayesian smoothing using Dirichlet priors.** The Dirichlet smoothing approach is similar to Jelinek–Mercer smoothing (Jelinek & Mercer, 1980) because it also uses a linear interpolation of the MLE model with another model. The model obtained by Dirichlet smoothing is given in the equation.

$$P(t|\theta_d) = \frac{|d|}{|d|+\mu} P_{MLE}(t|\theta_d) + \frac{\mu}{|d|+\mu} P_{MLE}(t|\theta_C).$$

In the equation, $P_{MLE}(t|\theta_C)$ is an MLE model constructed from a collection of documents $C$ to smooth the probability of the document model, $\mu$ is the interpolation weight, and $|d|$ is the length of document $d$. In our experiments, we will use the set of documents which arrive before document $d$ as Set C. In this smoothing approach, $\mu$ is obtained with training.

**LM-based ND: Shrinkage smoothing.** This smoothing approach assumes that each document is generated by the contribution of three LMs: a document model, a topic model, and a background model—in our case, a Turkish model (Allan, Wade, & Bolivar, 2003). Calculation of an LM with shrinkage smoothing is made as follows where $P_{MLE}(t|\theta_T)$ is the MLE model generated for the topic of document $d$ and $P_{MLE}(t|\theta_{TU})$ is the MLE model generated for Turkish.

$$P(t|\theta_d) = \lambda_d P_{MLE}(t|\theta_d) + \lambda_T P_{MLE}(t|\theta_T) + \lambda_{TU} P_{MLE}(t|\theta_{TU}).$$

Interpolation weights for the corresponding LM are shown as $\lambda_d$, $\lambda_T$, and $\lambda_{TU}$ where $\lambda_d + \lambda_T + \lambda_{TU} = 1$. These weights are obtained by training. In our experiments, $P_{MLE}(t|\theta_T)$ is generated by the topic description which is expanded by the first story of the topic. This is the maximum likelihood estimate of the probability made from a text that contains the topic description (which was provided by the annotators during construction of test collection *BilCol-2005*) and the first story of the topic to which the document belongs. Allan, Wade, and Bolivar (2003) also used TREC topic descriptions for topic models. The Turkish model $P_{MLE}(t|\theta_{TU})$ is generated by using a reference collection, the Milliyet Collection (Can, Kocberber, Balcik et al., 2008), which contains about 325,000 documents that are news from the *Milliyet* newspaper between Years 2001 and 2004. (Documents of Year 2005 of this collection are excluded to prevent any possible bias.) This corpus was utilized in other studies for information retrieval experiments (Can, Kocberber, Balcik et al., 2008) and again as a reference corpus for calculation of IDF statistics (Can et al., 2010).

*Adaptation of LMs to ND.* LMs previously have been used as novelty measures in different studies. In Allan et al. (2001), the occurrences of words in sentences are assumed to be independent from each other, and the probability of a sentence $s$ being generated by a model $\theta$ is calculated as in the following equation where $t$ represents terms and $s$ represents sentences. The root $|s|$ is taken for length normalization.

$$P(s|\theta) = \prod_{t \in s} P(t|\theta)^{\frac{1}{|s|}}.$$

Later, these values are directly used as novelty scores. This method seems to depend heavily on the quality of smoothing since one unrealistic (i.e., small) probability can make the result unreliable because of the multiplications. KL divergence is another measure used for utilizing LMs in ND (Allan, Wade, & Bolivar, 2003). KL divergence is used to find the distance between two probabilistic distributions. Calculation of KL divergence between two LMs, $\theta_1$ and $\theta_2$ is given the following equation.

$$KL(\theta_1, \theta_2) = \sum_t P(t|\theta_1) \log \frac{P(t|\theta_1)}{P(t|\theta_2)}.$$

As the formula suggests, KL divergence is an asymmetric measure where $KL(\theta_1, \theta_2)$ and $KL(\theta_2, \theta_1)$ do not necessarily


DOI: 10.1002/asi

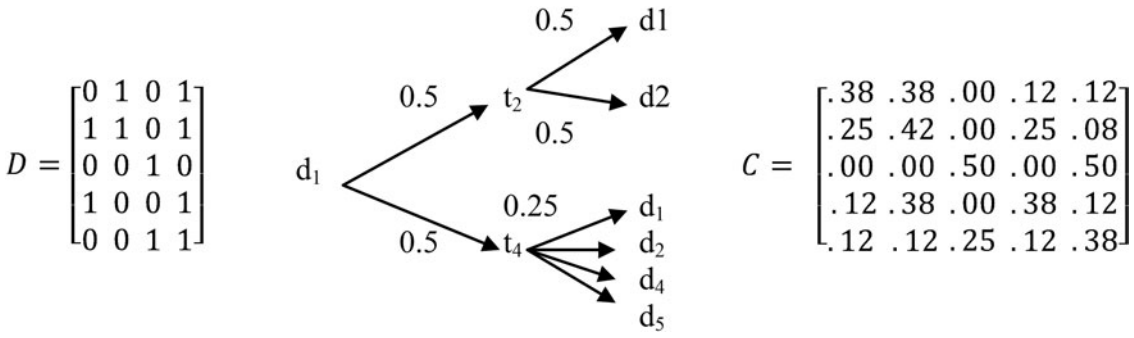


FIG. 5. Example transformation from the *D* matrix to the *C* matrix.

have the same values. This property makes it an appropriate measure for ND (Zhang et al., 2002).

In this study, we also use KL divergence as the novelty measure for LM-based ND. We follow the nonaggregate approach (discussed earlier); that is, for an incoming document, $d_t$, we calculate KL divergence between the document model and every previous-document's model. If KL divergence between $d_t$ and any of the previous documents is less than the threshold, $d_t$ is labeled as *not novel*. This comparison has a similar intuition as does the CS-based method (except that KL divergence is a distance measure, and thus a smaller value denotes higher resemblance).

*CC-based ND.* CC is a concept to quantify the extent to which a document is covered by another document (Can & Ozkarahan, 1990). The following equation shows the calculation of CC.

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

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

In the formula, $n$ and $m$ represent the number of terms and documents, respectively, in the document-term matrix $D$ of a set of documents, $d_{ik}$ is the number of occurrences of term $k$ in document $i$ where $1 \le i \le m$ and $1 \le k \le n$. Reciprocals of $i$th row sum and $k$th column sum of the $D$ matrix are represented as $\alpha_i$ and $\beta_k$, respectively.

Coverage of document $i$ by document $j$, $c_{ij} (1 \le i \le m, 1 \le j \le m)$, is the probability of selecting any term of document $i$ from document $j$. Calculation is done as a two-stage probability experiment. An illustration of the construction of the $C$ matrix is given in Figure 5, which is adapted from Can et al. (2010). The leftmost part shows an example document-term matrix which consists of five documents ($d_1$, $d_2$, $d_3$, $d_4$, $d_5$) and four terms ($t_1$, $t_2$, $t_3$, $t_4$). As stated in Can and Ozkarahan (1990), all documents should at least have one nonzero entry in the $D$ matrix, they should contain at least one term, and each term should at least be contained by one document. The $D$ matrix contains binary values in this example, but it also may be weighted. In the middle of Figure 5, an example of a double-stage probability experiment is given. In the first stage, a term is chosen randomly from $d_1$. Since the document has two terms, selection probabilities of both terms are 0.5 (obtained by $\alpha_1$). This stage is handled by the first part of the formula. In the second stage, the selected term is randomly chosen from a document. For example, if $t_4$ is considered, it may be selected from four documents with 0.25 probabilities (obtained by $\beta_4$). This stage is handled by the second part of the formula. The last part of the figure shows the constructed $C$ matrix, an $m \times m$ matrix, from the $D$ matrix which contains the $c_{ij}$ values.

**Motivation for using CC as a novelty measure.** The CC values are probabilities that show the characteristics of probabilistic observations. All $c_{ij}$ values vary between 0 and 1, with some restrictions (Can & Ozkarahan, 1990). If two documents contain no common terms, coverage of one by the other one is 0. The row sum of the $C$ matrix is equal to 1, which shows that the sum of probabilities of a document covered by itself and the other documents is equal to 1. A document's coverage of itself is called the *decoupling coefficient* and is shown by the $c_{ii}$ value for $1 \le i \le m$. If a document contains terms which only exist in it, the decoupling coefficient of the document is 1, and its coverage by all other documents is equal to 0.

The CC value is an asymmetric measure which can easily be shown by an example set of two documents in which one of the documents contains the other one. Coverage of the smaller document by the superset is greater than is the coverage of the superset by the subset. This asymmetric property makes the CC concept useful as a novelty measure because the same situation exists in ND. Consider two documents, $d_1$ and $d_2$ (see Figure 6), which may be regarded as tracking documents in a topic. Information contained in the documents is shown as A and B, where $d_1$ contains Information A and $d_2$ contains Information A and B. In the first case, $d_1$ arrives at $t_1$ and contains Information A, which was not delivered before. Thus, $d_1$ is novel. At time $t_2$, $d_2$ arrives and contains Information A and B. Information B was not reported before $t_2$, so this document also is labeled as *novel*. To observe the asymmetric property, we swap the order of the arrival of documents. In the swapped case, $d_2$ arrives at $t_1$ and is labeled as *novel* since it contains A and B, which were not given before. However, $d_1$, which arrives at $t_2$, contains no novel information since A already was given in $d_2$ before. This property may not be handled well by symmetric similarity measures such as CS since similarity between $d_1$ and $d_2$ is calculated


DOI: 10.1002/asi

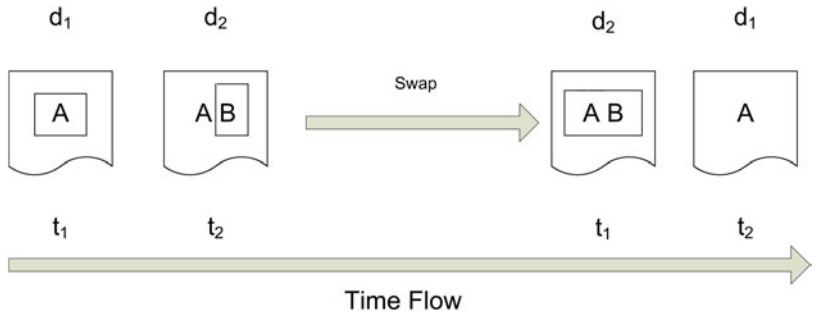


FIG. 6. Example case of asymmetry in novelty detection.

regardless of their arrival times. In CC, coverage of $d_1$ by $d_2$ is expected to be larger than the coverage of $d_2$ by $d_1$ in this specific case, which satisfies the ND property.

For deciding novelty, as in CS-based ND, we look for the condition that coverage of a document by all of the previous documents is below a threshold value.

## Experimental Evaluation

In this section, we first explain the evaluation measures used in this study and the preprocessing that we apply on texts. We then report the evaluation results of our methods and discuss them.

### *Evaluation Measures*

In TREC novelty tracks, the F-measure is used as the evaluation criterion (Harman, 2002; Soboroff, 2004; Soboroff & Harman, 2003). If we want to give equal weights to precision and recall, the F-measure can be calculated like the following:

$$F - measure = \frac{2 \cdot P \cdot R}{P + R},$$

where $P$ indicates precision and $R$ is recall.

For a topic, precision is defined as the ratio of number of correct novel documents identified by the system to the number of all documents identified by the system as novel. Recall is the ratio of correctly labeled novel documents by the system to the total novel documents. In this study, we use the macro-averaged F-measure, as in TREC novelty tracks.

Before proceeding with the methods, some preprocessing methods are applied on the texts, which are described next.

### *Preprocessing*

There are generally three steps of preprocessing applied on natural language texts: tokenization, stop-word elimination, and stemming. Tokenization, in this context, is the identification of the word boundaries. In most languages, including Turkish, tokenization is straightforward by tokenizing with respect to the spaces and punctuation marks.

Stop-words may affect performance of algorithms since they occur very frequently in texts. These words do not distinguish sentences/documents from each other; elimination of them is expected to increase system performance. In Turkish information retrieval, the effects of stop-word elimination are examined (Can, Kocberber, Balcik et al., 2008). The authors utilize three stop-word lists and report no significant difference between effectiveness of these different configurations. As a more similar study to ND, Can et al. (2010) showed that using a stop-word list significantly increases the effectiveness in new-event detection. However, there was no significant difference between the effectiveness of the system with the longest stop-word list and the system with a shorter list. In this work, we utilize the longest stop-word list, which contains 217 words taken from Kardaş (2009). This is a manually extended version of a shorter stop-word list (Can, Kocberber, Balcik et al., 2008). All letters are converted to lower case.

Different stemming algorithms are used to find the stems of the words so that word comparisons may be more reliable. In this work, a stemming heuristic called *Fixed Prefix Stemming* is utilized. Turkish is an agglutinative language in which suffixes are used to derive words with different meanings (Lewis, 1967). In fixed prefix stemming, a word's first $N$ characters are used as the word stem. For example, for the word *ekmekçi* (bread seller or bread maker), the first-five (F5) stem is *ekmek* (bread). Turkish's agglutinative property makes fixed prefix stemming an appropriate approach. Can, Kocberber, Balcik et al. (2008) showed that in information retrieval, fixed prefix stemming performs comparably with more sophisticated approaches such as a lemmatizer-based stemmer (Altintas, Can, & Patton, 2007). In addition, in new-event detection, it is shown that systems using F6 are one of the best performing ones (Can et al., 2010). In this study, we utilize F6 stemming as a result of this observation.

TABLE 2. Average results of random baseline.

| Ground truth | Precision | Recall | F-measure |
|---|---|---|---|
| Pessimistic | 0.498 | 0.500 | 0.491 |
| Optimistic | 0.678 | 0.500 | 0.573 |

## Results

### *Turkish ND Results*

*Random baseline system.* In Table 2, we present the results of the random baseline system. Note that the random baseline performs as expected. As stated earlier, for a challenging


DOI: 10.1002/asi

TABLE 3. Average results of the cosine similarity-based novelty detection method according to both ground truth data.

| | Training | | | Testing | | |
|---|---|---|---|---|---|---|
| Ground truth | Precision | Recall | F-measure | Precision | Recall | F-measure |
| Pessimistic | 0.630 | 0.935 | 0.741 | 0.631 | 0.923 | 0.738 |
| Optimistic | 0.778 | 0.963 | 0.857 | 0.776 | 0.954 | 0.852 |

TABLE 4. Results of the language model-based novelty detection method.

| | | Training | | | Testing | | |
|---|---|---|---|---|---|---|---|
| Smoothing approach | Ground truth | Precision | Recall | F-measure | Recall | Precision | F-measure |
| Dirichlet | Pessimistic | 0.747 | 0.904 | 0.806 | 0.741 | 0.900 | 0.801 |
| | Optimistic | 0.859 | 0.929 | 0.890 | 0.859 | 0.930 | 0.889 |
| Shrinkage | Pessimistic | 0.750 | 0.892 | 0.802 | 0.744 | 0.887 | 0.796 |
| | Optimistic | 0.841 | 0.942 | 0.885 | 0.838 | 0.933 | 0.880 |

test collection, random systems should not be able to perform well. In the pessimistic test collection, the performance of random baseline degrades since disagreement values are taken as not novel; that is, there appears to be less novel documents. In the following sections, we compare the results of the proposed methods with each other and with those of the random baseline.

*CS-based ND.* Results of the CS-based ND method according to both ground truth data are given in Table 3. Results show that this method significantly outperforms the baseline in terms of statistical tests, $p \ll 0.001$.

In this method, results according to optimistic ground truth data are higher because of the appropriateness of the method for a less strict novelty definition. Zhang et al. (2002) also had similar observations that their methods better modeled a less strict redundancy definition.

*LM-based ND.* Results of the LM-based ND method with two different smoothing approaches are given in Table 4. Shrinkage smoothing has more smoothing power and ideally has the ability to more accurately approximate probabilities, so we would expect Shrinkage to outperform Dirichlet smoothing in both ground truth type, but the algorithm produces similar results with both of the smoothing approaches (i.e., there is no statistically significant difference). The LM-based ND method also significantly outperforms the baseline in terms of statistical tests, $p \ll 0.001$.

Results are consistent with both Allan, Wade, & Bolivar (2003) and Zhang et al. (2002). In both of these studies, the Shrinkage and Dirichlet smoothing approaches have similar performance values.

*CC-based ND.* In this section, we provide the results of the CC-based ND method and compare it with the best configurations of the previously presented results (see Table 5). The best performing method, in terms of F-measure, is the LM-based ND with Dirichlet smoothing: It significantly outperforms the other two methods statistically, $p \ll 0.002$. This observation is generally consistent with ND studies conducted in English (Soboroff, 2004). Also as stated earlier, KL divergence is an appropriate measure for novelty because of its asymmetry. Smoothing is an important issue for LMs and Dirichlet smoothing seems to be successful in smoothing. In addition, it is easy to calculate and does not require any reference collection. The results with the Dirichlet smoothing approach show that the LM is highly successful; it provides a precision value of 0.859, a recall value of 0.930, and an F-measure value of 0.889 with the optimistic ground truth data; and can be used in real-life applications.

The second best performing system, CS-based ND is also one of the best performers in ND studies in English. This method is convenient to use because it does not require usage of a complex term-weighting function and generally works well with raw term frequencies (Allan et al., 2002).

CC as the least effective proposed method significantly outperforms random baseline in terms of statistical tests, $p \ll 0.001$, in both of the ground truth data. When compared to the LM method, the advantage of the CC-based ND method is that it only has one parameter.

*Effects of category-based threshold learning.* In this section, we report and compare the results of category-based threshold learning with general threshold learning. As can be seen in Table 6, there is no significant difference between the performances obtained by category-based threshold learning and general learning (see the $p$ values given in the last column). Although there is no significant difference, these results are promising; if there would be enough topics from every category, better results may be obtained by category-based learning. In this setup, since there are 59 topics and 11 categories, some categories have very few topics (e.g., 2). Even


DOI: 10.1002/asi

TABLE 5. Results of all methods' best configurations.

| | | Training | | | Testing | | |
|---|---|---|---|---|---|---|---|
| Method | Ground truth | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CC | Pessimistic | 0.550 | 0.928 | 0.681 | 0.542 | 0.923 | 0.672 |
| | Optimistic | 0.689 | 0.980 | 0.806 | 0.686 | 0.973 | 0.801 |
| LM Dirichlet | Pessimistic | 0.747 | 0.904 | 0.806 | 0.741 | 0.900 | 0.801 |
| | Optimistic | 0.859 | 0.929 | 0.890 | 0.859 | 0.930 | 0.889 |
| Cosine | Pessimistic | 0.630 | 0.935 | 0.741 | 0.631 | 0.923 | 0.738 |
| | Optimistic | 0.778 | 0.963 | 0.857 | 0.776 | 0.954 | 0.852 |
| Random | Pessimistic | No training results | | | 0.498 | 0.500 | 0.491 |
| | Optimistic | | | | 0.678 | 0.500 | 0.573 |

TABLE 6. Results of best performances of each system with general and category-based threshold learning.

| Method | Ground truth | General F-measure | Category F-measure | $p$ |
|---|---|---|---|---|
| CC | Pessimistic | 0.672 | 0.664 | 0.164 |
| | Optimistic | 0.801 | 0.798 | 0.677 |
| Cosine | Pessimistic | 0.738 | 0.732 | 0.625 |
| | Optimistic | 0.852 | 0.850 | 0.751 |
| LM Dirichlet | Pessimistic | 0.801 | 0.797 | 0.626 |
| | Optimistic | 0.889 | 0.887 | 0.409 |

TABLE 7. Parameter values for each novelty detection method on *BilNov-2005* learned and used in general threshold learning.

| | | Ground truth | |
|---|---|---|---|
| Method | Parameter | Pessimistic | Optimistic |
| Cosine | Similarity threshold | 0.79 | 0.89 |
| LM Dirichlet | KL threshold | 4.42 | 2.37 |
| | $\mu$ | 0.16 | 0.74 |
| LM Shrinkage | KL threshold | 3.95 | 1.97 |
| | $\lambda_d$ | 0.89 | 0.89 |
| | $\lambda_T$ | 0.10 | 0.10 |
| | $\lambda_{TU}$ | 0.01 | 0.01 |
| CC | Cover threshold | 0.21 | 0.32 |

if we apply leave-one-out cross-validation, the data size may still be insufficient to accurately learn a threshold value. Categories (or broader topics) are studied in FSD and also in a TREC novelty track as event and opinion types, but this type of category information has not been utilized. These results show that category information usage deserves further attention. These results also provide evidence about the robustness of the methods.

*Method parameters.* From the pragmatic perspective, the values of the parameters are interesting. As described earlier, we have two different approaches to optimize the method parameters: general threshold learning and category-based threshold learning. General threshold learning is 30-fold cross-validation applied over all topics. In $k$-fold ($k = 30$ in our general threshold learning scheme) cross-validation, data are divided into $k$ folds. Then, training and testing are repeated $k$ times with different $k - 1$ of the folds being used as the training set, and the remaining 1 fold as the testing set. At each repetition, parameter values are learned from the training set and applied on the testing set. For each repetition, since the training sets are different, learned parameter values may vary. In Tables 7 and 8, we present the parameter values that are the learned parameter values for the highest number of the repetitions; for example, if a value is optimal for a parameter in 20 of 30 repetitions, it is reported.

Table 7 lists the parameter values learned by general threshold learning and used in the test phase. Explanations of the parameters are given in the corresponding ND methods. As expected, the similarity measure-based (i.e., CS- and CC-based) parameters are estimated to be lower with pessimistic ground truth than with the optimistic ground truth. This is because in the pessimistic ground truth, the number of novel labeled documents is less than that of the optimistic one. Thus, it is reasonable for systems to lower similarity thresholds to make labeling a document novel more challenging. KL thresholds in LMs are distance measures, so they are higher in the pessimistic case. When we consider $\mu$ in LM Dirichlet, we see that the effect of smoothing is more powerful with the optimistic ground truth because the value of $\mu$ is higher. In LM Shrinkage, smoothing with the reference collection seems to have a small effect since it has a small weight ($\lambda_{TU}$).

Table 8 presents the parameter values learned and used in category-based threshold learning (there is no column for LM Shrinkage because there were no experiments conducted on LM Shrinkage in category-based threshold learning). In category-based threshold learning, we applied leave-one-out cross-validation on topics from the same category instead of using all topics together. Leave-one-out cross-validation is the special-case cross-validation where number of folds is equal to the data size. Since parameter values are learned specific to the categories, we report values separately for each category. Ordering of categories in terms of strictness


DOI: 10.1002/asi

TABLE 8. Parameter values for each novelty detection method on *BilNov-2005* learned and used in category-based threshold learning.

| | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Cosine | | LM Dirichlet | | | | Cover coefficient | |
| | Pessimistic | Optimistic | Pessimistic | | Optimistic | | Pessimistic | Optimistic |
| Category | Similarity threshold | Similarity threshold | KL threshold | $\mu$ | KL threshold | $\mu$ | Cover threshold | Cover threshold |
| Scandals/Hearings | 0.79 | 0.89 | 3.32 | 0.58 | 2.21 | 0.95 | 0.37 | 0.37 |
| Legal/Criminal Cases | 0.74 | 0.79 | 4.42 | 0.16 | 2.58 | 0.26 | 0.16 | 0.21 |
| Accidents | 0.79 | 0.79 | 3.68 | 0.84 | 2.58 | 0.16 | 0.21 | 0.42 |
| Acts of Violence or War | 0.79 | 0.89 | 6.63 | 0.11 | 1.84 | 0.05 | 0.21 | 0.53 |
| Science and Discovery News | 0.68 | 0.84 | 4.42 | 0.26 | 2.95 | 0.63 | 0.16 | 0.26 |
| Financial News | 0.84 | 0.95 | 3.68 | 0.95 | 1.84 | 0.05 | 0.11 | 0.42 |
| News Laws | 0.84 | 0.84 | 3.32 | 0.32 | 2.58 | 0.21 | 0.21 | 0.47 |
| Sports News | 0.74 | 0.84 | 1.84 | 0.53 | 1.84 | 0.11 | 0.16 | 0.16 |
| Political and Diplomatic Meetings | 0.68 | 0.89 | 5.53 | 0.42 | 4.05 | 0.58 | 0.11 | 0.16 |
| Celebrity/Human Interest News | 0.74 | 0.79 | 4.05 | 0.32 | 2.21 | 0.47 | 0.21 | 0.26 |
| Miscellaneous News | 0.68 | 0.79 | 4.79 | 0.68 | 2.95 | 0.11 | 0.16 | 0.26 |

TABLE 9. Test results for cover coefficient-based novelty detection method and 5 participants of TREC 2004.

| Participant (Run Name) | Precision | Recall | F-measure |
|---|---|---|---|
| Dublin City University (CDVP4nterf1) | 0.4904 | 0.9038 | 0.6217 |
| Meiji University (MeijiHIL2WRS) | 0.4790 | 0.9310 | 0.6188 |
| University of Massachussetts, Amherst (CIIRT2R2) | 0.4712 | 0.9544 | 0.6176 |
| 31 omitted results | | | |
| Center for Computing Sciences (ccsmmr5t2) | 0.4326 | 0.9938 | 0.5880 |
| Cover coefficient | 0.4334 | 1.0000 | 0.5867 |
| Meiji University (MeijiHIL2CS) | 0.4246 | 0.9952 | 0.5797 |
| 18 omitted results | | | |

of novelty definition for different methods is not highly correlated. Even the ordering in the same method differs for different ground truth types. For example, the "Political and Diplomatic Meetings" category has the smallest CS threshold value in terms of pessimistic ground truth, but not for optimistic ground truth. As mentioned earlier, a smaller similarity threshold means a stricter novelty definition. (Reversely, smaller distance measure, KL, means a less strict novelty definition.) Because of the low correlation between methods, it is hard to make an accurate ordering of the categories in terms of strictness of novelty definition. But it is reasonable to assume that if we have more topics per category, we would be able to examine some patterns.

### *TREC Novelty Track 2004 Results*

We also experimented with the TREC 2004 test collection to see effects of applying the same method on test collections in different languages. We used TREC Novelty 2003 data for training and 2004 data for testing (TREC, 2011). We only ran the CC-based ND method on TREC 2004 data since both CS and LMs were used in the track by other participants. The results we provide are for Task 2, which is finding novel sentences when relevant sentences are given, because relevant sentence detection is beyond the scope of our work.

The results can be seen in Table 9. There were 55 participants. We only included the results of five runs from Task 2 to reflect the performance figures obtained. The first three rows show the best performing three systems of Task 2. The important result for our comparison purposes is *CIIRT2R2* because they used CS for ND (Jaleel et al., 2004). This finding is similar to our findings in *BilNov-2005* that the CS-based ND method outperforms the CC-based method. In addition, in their previous study, Allan, Wade, and Bolivar (2003) showed that LM-based ND methods outperform the CS-based method in the TREC 2003 data. When all of these results are examined, we can arguably claim that results are consistent with the results in Turkish.

The CC-based ND outperforms the baseline in Task 2 and is ranked 36th out of 55 participants. We are optimistic that its performance can be improved by further research. For example, some further adaptations may boost performance of the method, such as a normalization factor to prevent possible anomalies caused by the differences in lengths of


DOI: 10.1002/asi

sentences. In addition, a complex threshold mechanism can be employed.

## Conclusions and Future Work

This work contributes to research on ND in TT; to the best of our knowledge, it is the first large-scale ND study in TT in literature. One major goal of this study is to construct a reliable ND test collection that serves as a ground truth and can be used in the development and evaluation of ND algorithms for TT. For this purpose, we built the *BilNov-2005* ND test collection, which was constructed from the topics of the *BilCol-2005* (Can et al., 2010). *BilNov-2005* is available to other researchers (BilNov-2005, 2010). For quality assessment of the test collection, we consider the topic lengths, novelty ratios, and interannotator agreements.

Using *BilNov-2005*, we present pioneering benchmark findings on ND for TT in Turkish. For this purpose, we examine three ND methods: a CS-based method, an LM-based method, and a CC-based method. The first two methods were motivated by previous studies on ND. For the LM-based ND method, we show that a simpler smoothing approach, Dirichlet smoothing, provides a performance similar to a more complex smoothing approach, Shrinkage smoothing. In addition to these two methods, we propose a CC-based ND method. By following the tradition of TDT studies, we establish a baseline that shows the performance of random decisions for ND. For the first time in ND, we consider a category-based threshold learning method, which uses topics from the same category when learning a threshold. It is motivated by differences between characteristics of news from different categories. Although the results of category-based and general threshold learning do not have any significant differences, it is promising to see that even with a small set of topics from the same category, learning can be conducted without decreasing performance. Finally, we provide the results of a CC-based ND method in the TREC 2004 Novelty Track test collection; it is ranked 36th out of 55 participants.

Although ND was studied in information retrieval for 3 years in TREC novelty tracks (Harman, 2002; Soboroff, 2004; Soboroff & Harman, 2003), there is still much work in both information retrieval and other domains. Most of the ND methods are domain independent and can work with any set of documents. ND in patient reports, intelligence applications, blog and web mining, and information filtering are some other possible application areas. Our results show that in ND for TT, the LM is highly successful and can be used in real-life applications. Some future research possibilities for ND studies, among others, include the following directions. Category information can be utilized in a more sophisticated way and evaluated with a larger test collection containing several topics per category. When working on documents, instead of considering documents as a whole, sentences may be processed separately. In such environments, some of the sentences in a document can be irrelevant and may contain novel information. Such sentences may be eliminated before ND. For an evaluation of sentence-level relevance detection, a TREC novelty track test collection may be used or a new test collection may be created.

## Acknowledgments

We thank our colleagues, friends, and students for their topic annotations. We are grateful to the anonymous referees for their constructive comments and suggestions. We also thank Gönenç Ercan and Sengör Altıngövde for their valuable comments on an earlier draft, and Çağdaş Öcalan for his support of the project. This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) Grant 108E074. 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

Automatic Content Extraction Workshops. (2005). The ACE 2005 (ACE05) Evaluation Plan. Retrieved from http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf

Aksoy, C. (2010). Novelty detection in topic tracking (Unpublished master's thesis). Bilkent University, Department of Computer Engineering, Bilkent, Ankara, Turkey. Retrieved from http://www.cs.bilkent.edu.tr/~canf/bilir_web/theses/cemAksoyThesis.pdf

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., …Zhai, C. (2003). Challenges in information retrieval and language modeling: Report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, 2003. SIGIR Forum, 37(1), 31–47.

Allan, J., Carbonell, J., Doddington, G., & Yamron, J. (1998). Topic detection and tracking pilot study final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (pp. 194–218). San Francisco: Kaufmann.

Allan, J., Gupta, R., & Khandelwal, V. (2001). Temporal summaries of new topics. In Proceedings of the 24th International Conference on Research and Development in Information Retrieval (ACM SIGIR '01) (pp. 10–18). 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., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. In Proceedings of the 26th International Conference on Research and Development in Information Retrieval (ACM SIGIR '03) (pp. 314–321). New York: ACM Press.

Altintas, K., Can, F., & Patton, J.M. (2007). Language change quantification using time-separated parallel translations. Literary & Linguistic Computing, 22(4), 375–393.

BilNov-2005. (2010). Bilkent novelty detection test collection. Retrieved from http://www.cs.bilkent.edu.tr/~canf/BilNov2005/BilNov2005%20Info.htm

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., & Uyar, E. (2010). New event detection and topic tracking in Turkish. Journal of the American Society for Information Science and Technology, 61(4), 802–819.

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

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., & Uyar, E. (2008). Bilkent News Portal: A personalizable system with new event


DOI: 10.1002/asi

detection and tracking capabilities. In Proceedings of the 31st International Conference on Research and Development in Information Retrieval (ACM SIGIR '08) (p. 885). New York: ACM Press.

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.

Cheng, T. (2005). Apply novelty detection on topic detection and tracking (Unpublished master's thesis). Institute of Information Management, National Yunlin University of Science & Technology, Yunlin, Taiwan. Retrieved from http://ethesys.yuntech.edu.tw/ETD-db/ETD-search/view_etd?URN=etd-0719105-130254

Chowdhury, A., Frieder, O., Grossman, D.A., & McCabe, M.C. (2002). Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems, 20(2), 171–191.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Conrad, J.G., & Schriber, C.P. (2006). Managing déjà vu: Collection building for the identification of nonidentical duplicate documents. Journal of the American Society for Information Science and Technology, 57(7), 921–932.

Dang, H.T., & Owczarzak, K. (2008). Overview of the TAC 2008 update summarization task. In Proceedings of the Text Analysis Conference. Retrieved from http://www.nist.gov/tac/publications/2008/additional. papers/update_summ_overview08.proceedings.pdf

Document Understanding Conference. (2007). DUC 2007: Task, documents, and measures. Retrieved from http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html

The Economist. (2011, July 7). Special report: The news industry. Retrieved from http://www.economist.com/node/18904136

Eichmann, D., Zhang, Y., Bradshaw, S., Qiu, X.Y., Zhou, L., Srinivasan, P., ... Wong, H. (2004). Novelty, question answering and genomics: The University of Iowa response. In Proceedings of the 13th Text Retrieval Conference. Retrieved from http://comminfo.rutgers.edu/~muresan/IR/TREC/Proceedings/t13_proceedings/papers/uiowa.novelty.qa.geo.pdf

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.

Harman, D. (2002). Overview of the TREC 2002 novelty track. In Proceedings of the 11th Text Retrieval Conference. Retrieved from http://trec.nist.gov/pubs/trec11/papers/NOVELTY.OVER.pdf

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85–126.

Jain, A.K., & Dubes, R.C. (1988). Algorithms for clustering data. Upper Saddle River, NJ: Prentice Hall.

Jaleel, N.A., Allan, J., Croft, W.B., Diaz, F., Larkey, L.S., Li, X., Smucker, M.D., & Wade, C. (2004). UMass at TREC 2004: Novelty and HARD. In Proceedings of the 13th Text Retrieval Conference. Retrieved from http://comminfo.rutgers.edu/~muresan/IR/TREC/Proceedings/t13_proceedings/papers/umass.novelty.hard.pdf

Jelinek, F., & Mercer, R.L. (1980). Interpolated estimation of Markov source parameters from sparse data. In Proceedings of the Workshop on Pattern Recognition in Practice. Amsterdam, The Netherlands: North-Holland.

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

Kumaran, G., & Allan, J. (2004). Text classification and named entities for new event detection. In Proceedings of the 27th International Conference on Research and Development in Information Retrieval (ACM SIGIR '04) (pp. 297–304). New York: ACM Press.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Larkey, L.S., Allan, J., Connell, M.E., Bolivar, A., & Wade, C. (2002). UMass at TREC 2002: Cross language and novelty tracks. In Proceedings of the 11th Text Retrieval Conference. Retrieved from http://trec.nist.gov/pubs/trec11/papers/umass.wade.pdf

Lewis, G.L. (1967). Turkish grammar. Oxford, United Kingdom: Clarendon Press.

Li, X., & Croft, W.B. (2008). An information-pattern-based approach to novelty detection. Information Processing & Management, 44(3), 1159–1188.

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

Markou, M., & Singh, S. (2003). Novelty detection: A review—Part 1: Statistical approaches. Signal Processing, 83(12), 2481–2497.

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

Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In Proceedings of the 11th International Conference on Knowledge Discovery in Data Mining (ACM SIGKDD '05) (pp. 198–207). New York: ACM Press.

Moon, T.K. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6), 47–60.

Öcalan, H.Ç. (2009). Bilkent News Portal: A system with new event detection and tracking capabilities (Unpublished master thesis). Bilkent University, Computer Engineering Department, Bilkent, Ankara, Turkey. Retrieved 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, Amherst, Department of Computer Science. Retrieved from http://ciir.cs.umass.edu/pubfiles/ir-179.pdf

Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval (ACM SIGIR '98) (pp. 275–281). New York: ACM Press.

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

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

Soboroff, I. (2004). Overview of the TREC 2004 novelty track. In Proceedings of the 13th Text Retrieval Conference. Retrieved from http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf

Soboroff, I., & Harman, D. (2003). Overview of the TREC 2003 novelty track. In Proceedings of the 20th Text Retrieval Conference. Retrieved from http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf

Soboroff, I., & Harman, D. (2005). Novelty detection: The TREC experience. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05) (pp. 105–112). Morristown, NJ: Association for Computational Linguistics.

Sweeney, S.O., Crestani, F., & Losada, D.E. (2008). Show me more: Incremental length summarisation using novelty detection. Information Processing & Management, 44(2), 663–686.

Text Analysis Conference. (2009). Overview of TAC 2009 summarization track. Retrieved from http://www.nist.gov/tac/publications/2009/presentations/TAC2009_Summ_overview.pdf

Topic Detection and Tracking Initiative. (2004). Annotation manual: Version 1.2. Retrieved from http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf

Text Retrieval Conference. (2011). TREC Results. Retrieved from http://trec.nist.gov/results/ (requires password).

Tsai, F.S., Tang, W., & Chan, K.L. (2010). Evaluation of novelty metrics for sentence-level novelty mining. Information Sciences, 180(12), 2359–2374.

Tsai, M.-F., Hsu, M.-H., & Chen, H.-H. (2004). Similarity computation in novelty detection and biomedical text categorization. In Proceedings of the 13th Text Retrieval Conference. Retrieved from http://comminfo.rutgers.edu/~muresan/IR/TREC/Proceedings/t13_ proceedings/papers/ntu.novelty.pdf

van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.
Varol, E., Can, F., Aykanat, C., & Kaya, O. (2011). CoDet: Sentence-based containment detection in news corpora. In Proceedings of the 20th Conference on Information and Knowledge Management (ACM CIKM '11) (pp. 2049–2052). New York: ACM Press.
Yang, Y., Pierce, T., & Carbonell, J. (1998). A study of retrospective and online event detection. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval (ACM SIGIR '98) (pp. 28–36). New York: ACM Press.
Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). Topic-conditioned novelty detection. In Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD '02) (pp. 688–693). New York: ACM Press.
Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22(2), 179–214.
Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '02) (pp. 81–88). New York: ACM Press.

## Appendix

Table A1. *BilNov-2005* Topic Information.[a]

| Topic no. and topic short description in Turkish–English (*BilCol-2005* Topic no.) | Topic category<br>Start date–end date (mm/dd) | No. of track documents | Novelty ratio: Pessimistic (%) | Novelty ratio: Optimistic (%) |
|---|---|---|---|---|
| 1. Kars'da trafik kazası 7 ölü 35 yaralı–Accident in Kars kills 7 injures 37 (1) | Accidents<br>05/28–12/16 | 20 | 45.00 | 70.00 |
| 2. Onur Air'in Avrupa'nın bazı ülkelerinde iniş kalkışının yasaklanması – Some Europian countries ban Onur Air flights (2) | Legal/criminal cases<br>05/12–05/17[b] | 80 | 60.00 | 62.50 |
| 3. Nema karşılığı kredi–Advanced payment based on dividends (4) | Financial news<br>02/08–11/14 | 31 | 64.52 | 80.65 |
| 4. Londra metrosunda patlama–London underground explosion (6) | Acts of violence or war<br>07/07–07/07 | 80 | 26.25 | 60.00 |
| 5. Çocuk tacizi skandalı – Child abuse scandal (7) | Scandals/hearings<br>01/26–03/09 | 80 | 56.25 | 78.75 |
| 6. Formula G–Formula G (8) | Sports news<br>07/04–08/30 | 20 | 60.00 | 80.00 |
| 7. Şemdinli olayları–Şemdinli events (11) | Scandals/hearings<br>11/9–11/12 | 80 | 59.49 | 73.42 |
| 8. Türkiye'de kuş gribi–Bird flu in Turkey (12) | Miscellaneous news<br>10/10–10/14 | 80 | 37.50 | 70.00 |
| 9. Fenerbahçe'nin şampiyon olması – Championship of Fenerbahçe (13) | Sports news<br>05/22–05/30 | 80 | 61.25 | 70.00 |
| 10. Mortgage Türkiye'de–Mortgage in Turkey (14) | New laws<br>01/07–06/13 | 80 | 55.00 | 71.25 |
| 11. 2005 Avrupa Basketbol şampiyonası –2005 European Basketball championship (15) | Sports news<br>01/15–11/07 | 78 | 43.59 | 64.10 |
| 12. Van Yüzüncü Yıl Üniversitesi rektörü Prof. Dr. Yücel Aşkın'ın tutuklanması–Arrest of Van Yüzüncü Yıl University's president Prof. Dr. Yücel Aşkın (16) | Scandals and hearings<br>10/14–10/22 | 80 | 55.00 | 65.00 |
| 13. Kral Fahd'ın hastaneye kaldırılması–King Fahd's hospitalization (17) | Celebrity/human interest news<br>05/27–08/11 | 51 | 56.86 | 76.47 |
| 14. Memurların bir üst dereceye çıkması–Promotion of government officers to a higher rank (18) | New laws<br>01/06–04/25 | 52 | 44.23 | 55.77 |
| 15. Bill Gates'in Türkiye'ye gelmesi–Bill Gates visits Turkey (19) | Celebrity/human interest news<br>01/30–02/06 | 17 | 70.59 | 76.47 |
| 16. Mısır'da üst üste patlamalar–Successive explosions in Egypt (20) | Acts of violence or war<br>07/23–07/26 | 80 | 37.50 | 63.75 |
| 17. Atillâ İlhan'ın vefat etmesi–Atillâ İlhan dies (21) | Celebrity/human interest news<br>10/11–12/19 | 40 | 52.50 | 80.00 |
| 18. Ata Türk'ün öldürülmesi–Murder of Ata Türk (22) | Legal/criminal cases<br>09/18–11/03 | 43 | 55.81 | 58.14 |
| 19. DT Genel Müdürü Lemi Bilgin'in görevden alınması–State theater general director Lemi Bilgin is taken from his post (23) | Celebrity/human interest news<br>08/19–12/05 | 63 | 69.84 | 80.95 |
| 20. Universiade 2005–Universiade 2005 (24) | Sport news<br>03/04–08/12 | 80 | 82.50 | 87.50 |
| 21. Yahya Murat Demirel'in Bulgaristan'da yakalanması–Capture of Yahya Murat Demirel in Bulgaria (25) | Legal/criminal cases<br>01/03–01/08 | 80 | 45.00 | 66.25 |

(*Continued*)


DOI: 10.1002/asi

Table A1. Continued.

| Topic no. and topic short description in Turkish–English (*BilCol-2005* Topic no.) | Topic category Start date–end date (mm/dd) | No. of track documents | Novelty ratio: Pessimistic (%) | Novelty ratio: Optimistic (%) |
|---|---|---|---|---|
| 22. Bağdat El-Ayma köprüsü üzerinde izdihamda çok sayıda insanın ölmesi–Stampede on Baghdat El-Ayma bridge kills many people (26) | Acts of violence or war 08/31–09/08 | 29 | 37.93 | 62.07 |
| 23. Prof. Dr. Sadettin Güner ve oğlunun Trabzon'da öldürülmesi–Murder of Prof. Dr. Sadettin Güner and his son in Trabzon (27) | Legal/criminal cases 01/08–10/25 | 41 | 56.10 | 68.29 |
| 24. Nermin Erbakan'ın tedavi altına alınması–Nermin Erbakan is under treatment (29) | Celebrity/human interest news 10/20–12/04 | 45 | 48.89 | 66.67 |
| 25. 15. Akdeniz Oyunları–Mediterranean Games (31) | Sport news 05/02–06/28 | 80 | 72.50 | 73.75 |
| 26. Kemal Derviş'in UNDP Başkanı seçilmesi ve göreve başlaması–Kemal Derviş is elected and started as head of UNDP (32) | Finance news 03/11–05/05 | 80 | 35.00 | 55.00 |
| 27. Caferi'nin tarihi Tahran ziyareti–Caferi's historical Tehran visit (33) | Political and diplomatic meetings 07/05–10/06 | 22 | 68.18 | 77.27 |
| 28. Gediz'de grizu patlaması–Mine gas explosion in Gediz (34) | Accidents 04/21–05/26 | 39 | 56.41 | 61.54 |
| 29. Sarıgül'ün kendini savunması–Sarıgül defends himself (35) | Political and diplomatic meetings 01/02–03/18 | 80 | 41.25 | 68.75 |
| 30. Paris'de göstericilerin polisle çatışması–Clash between police and demonstrators in Paris (36) | Acts of violence or war 10/29–11/07 | 80 | 42.50 | 72.50 |
| 31. 2005 Nobel Tıp ödülü gastrit ve ülserin bakterilerden kaynaklanması–Medical Nobel awarded for ulcer and gastritis study (39) | Science and discovery news 10/03–12/16 | 19 | 42.11 | 57.89 |
| 32. Kayseri Erciyes üniversitesi bebek ölümleri–Baby deaths at Kayseri Erciyes University (40) | Scandals/hearings 08/03–10/01 | 39 | 53.85 | 64.10 |
| 33. Marburg virüsünden ölenler–Marburg virus deaths (41) | Miscellaneous news 03/16–05/19 | 25 | 56.00 | 72.00 |
| 34. Gamze Özçelik'in görüntülerinin internette yayınlanması–Gamze Özçelik videos appear on the Internet (42) | Celebrity/human interest news 08/29–12/22 | 43 | 60.47 | 72.09 |
| 35. Türkiye'nin ilk yediz bebekleri–Turkey's first septuplets (43) | Science and discovery news 02/17–12/14 | 56 | 57.14 | 73.21 |
| 36. Yeni Türk ceza kanununun yürürlüğe girmesi–New Turkish criminal law goes into effect (44) | New laws 06/01–12/10 | 53 | 60.38 | 67.92 |
| 37. Saddam Hüseyin'in yargılanmaya başlanması–Trial of Saddam Hussein starts (45) | Legal/criminal cases 10/19–11/28 | 80 | 52.50 | 57.50 |
| 38. Beylikdüzü'nde çöpte patlama–Explosion in garbage in Beylikdüzü (46) | Acts of violence or war 11/18–11/22 | 17 | 47.06 | 58.82 |
| 39. Endonezya'nın Bali Adası'nda eşzamanlı patlamalar–Indenosia Bali Island concurrent bombings (47) | Acts of violence or war 10/01–10/04 | 15 | 33.33 | 60.00 |
| 40. Sahte rakı–Counterfreit rakı (48) | Legal/criminal cases 03/01–03/03 | 80 | 43.75 | 57.50 |
| 41. Hindistan'da bir saldırıda 66 kişi öldü–In India an attack kills 66 people (49) | Acts of violence or war 10/29–11/02 | 21 | 71.43 | 85.71 |
| 42. Bülent Ersoy ve Deniz Baykal polemiği–Polemic between [singer] Bülent Ersoy and [politician] Deniz Baykal (50) | Celebrity/human interest news 08/19–12/28 | 52 | 44.23 | 59.62 |
| 43. Sochi seferini yapan Ufuk-1 gemisinin yanması–Ufuk-1 ship on fire while sailing to Sochi (52) | Accidents 08/25–08/27 | 20 | 45.00 | 70.00 |
| 44. İstanbul'da Dünya Kadınlar Günü için gösteri yapanları coplayan üç polisin açığa alınması–Three policemen lay off after bludgeoning demonsters during World Women's Day (54) | Legal criminal cases 03/06–03/16 | 80 | 42.50 | 66.25 |
| 45. Kuşadası'nda minibüsdeki patlamada beş kişinin ölmesi–Five die in an explosion in a minibus in Kuşadası (55) | Acts of violence or war 07/16–07/19 | 50 | 28.00 | 54.00 |
| 46. Esenboğa Havalimanı iç hatlar terminali'nin yanması–Fire in the Esenboğa airport domestic terminal (56) | Accidents 11/14–12/19 | 18 | 38.89 | 72.22 |
| 47. Zeytinburnu'nda bir evde meydana gelen patlamada iki kişinin ölmesi–Two die in an explosion in a house in Zeytinburnu (57) | Acts of violence or war 08/08–08/11 | 28 | 32.14 | 57.14 |

(*Continued*)


DOI: 10.1002/asi

Table A1. Continued.

| Topic no. and topic short description in Turkish–English (*BilCol-2005* Topic no.) | Topic category Start date–end date (mm/dd) | No. of track documents | Novelty ratio: Pessimistic (%) | Novelty ratio: Optimistic (%) |
|---|---|---|---|---|
| 48. Malatya çocuk yuvası'nda işkence–Torture in Malatya kindergarten (58) | Scandals/hearings 10/26–10/28 | 80 | 56.25 | 81.25 |
| 49. Prof Dr. Kalaycı'nın silahlı saldırı sonucu öldürülmesi–Murder of Prof Dr. Kalaycı in an armed attack (60) | Legal/criminal cases 11/11–12/03 | 44 | 40.91 | 56.82 |
| 50. 15 yeni üniversite kuruluyor–15 new universities established (62) | New laws 11/12–12/31 | 59 | 33.90 | 59.32 |
| 51. Gaziantep'te tanker patlaması–Tanker explosion in Gaziantep (63) | Accidents 08/6–08/12 | 33 | 51.52 | 60.61 |
| 52. Kâzım Koyuncu'nun ölümü–Kâzım Koyuncu dies (66) | Celebrity/human interest news 06/25–10/31 | 30 | 66.67 | 73.33 |
| 53. Melih Kibar'ın ölümü–Melih Kibar dies (67) | Celebrity/human interest news 04/07–08/04 | 16 | 56.25 | 81.25 |
| 54. Japonya Osaka'da tren kazası–Train accident in Osaka, Japan (71) | Accidents 04/25–04/28 | 29 | 51.72 | 68.97 |
| 55. Yunanistan'da Türk bayrağına çirkin saldırı–Vandalism against Turkish flag in Greece (74) | Scandals/hearings 04/16–06/25 | 55 | 27.27 | 52.73 |
| 56. Maslak'ta patlama–Explosion in Maslak (75) | Acts of violence or war 10/15–11/01 | 30 | 40.00 | 73.33 |
| 57. Rum yolcu uçağının düşmesi–Cypriot passanger plane crashes (77) | Accidents 8/14–8/18 | 80 | 47.50 | 65.00 |
| 58. Zeytinburnu'nda geminin batması–Ship sinks in Zeytinburnu (79) | Accidents 03/13–03/15 | 38 | 28.95 | 60.53 |
| 59. İngiltere'de Osmanlı kültürü hakkında sergi açıldı–Ottoman culture exhibition opens in England (80) | Miscellaneous news 01/01–04/13 | 22 | 36.36 | 63.64 |
| Average | n/a | 50.89 | 49.89 | 67.79 |

[a]In this table, we provide topic information for the *BilNov-2005* test collection (Aksoy, 2010; *BilNov-2005*, 2010). It is based on the Topic Detection and Tracking test collection *BilCol-2005* (Can et al., 2010). *BilNov-2005* can be obtained by visiting the URL given in the related reference (BilNov-2005, 2010). The news categories are the same as defined for the Topic Detection and Tracking Initiative (2004) studies. In the following list, after each news category, the number of topics in that category is given within square brackets (e.g., there is no topic in the Elections category): elections [0], scandals/hearings [6], legal/criminal cases [8], natural disasters [0], accidents [8], acts of violence or war [10], science and discovery news [2], financial news [2], new laws [4], sports news [5], political and diplomatic meetings [2], celebrity/human interest news [9], and miscellaneous news [3].

[b]Note that the end date indicates the date of the 80th tracking news (not necessarily the end of the event); it is the same for other topics with 80 tracking news.


DOI: 10.1002/asi