NOVELTY DETECTION IN TOPIC TRACKING

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Cem Aksoy
July, 2010

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fazlı Can(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Seyit Koçberber (Co-Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Pınar Duygulu Şahin

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Dr. İlyas Çiçekli

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Nesim K. Erkip

Approved for the Institute of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Institute

ABSTRACT

NOVELTY DETECTION IN TOPIC TRACKING

Cem Aksoy
M.S. in Computer Engineering
Supervisors
Prof. Dr. Fazlı Can
Asst. Prof. Dr. Seyit Koçberber
July, 2010

News portals provide many services to the news consumers such as information retrieval, personalized information filtering, summarization and news clustering. Additionally, many news portals using multiple sources enable their users to evaluate developments from different perspectives by richening the content. However, increasing number of sources and incoming news makes it difficult for news consumers to find news of their interest in news portals. Different types of organizational operations are applied to ease browsing over the news for this reason. New event detection and tracking (NEDT) is one of these operations which aims to organize news with respect to the events that they report. NEDT may not also be enough by itself to satisfy the news consumers' needs because of the repetitions of information that may occur in the tracking news of a topic due to usage of multiple sources. In this thesis, we investigate usage of novelty detection (ND) in tracking news of a topic. For this aim, we built a Turkish ND experimental collection, BilNov, consisting of 59 topics with an average of 51 tracking news. We propose usage of three methods; cosine similarity-based ND method, language model-based ND method and cover coefficient-based ND method. Additionally, we experiment on category-based threshold learning which has not been worked on previously in ND literature. We also provide some experimental pointers for ND in Turkish such as restriction of document vector lengths and smoothing methods. Finally, we experiment on TREC Novelty Track 2004 dataset. Experiments conducted by using BilNov show that language model-based ND method outperforms other two methods significantly and category-based threshold learning has promising results when compared to general threshold learning.

Keywords: Novelty Detection, Topic Tracking.

ÖZET

KONU İZLEMEDE YENİLİK BULMA

Cem Aksoy
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticileri
Prof. Dr. Fazlı Can
Asst. Prof. Dr. Seyit Koçberber
Temmuz, 2010

Haber portalları okuyuculara bilgi erişimi, kişiselleştirilmiş bilgi filtreleme, özet çıkarma ve haber kümeleme gibi bir çok hizmet sunmaktadır. Bunlara ek olarak, pek çok haber portalı çok sayıda kaynaktan beslenerek kullanıcılarının gelişmeleri değişik açılardan değerlendirebilmelerini sağlamaktadır. Fakat artan haber kaynağı ve haber sayısı, haber okuyucularının kendi ilgi alanlarında olan haberleri bulabilmelerini zorlaştırmaktadır. Haberlerin kolay bir şekilde taranabilmesi için değişik düzenlemelerde bulunulmaktadır. Bu düzenlemelerden biri olan yeni olay bulma ve izleme (YOBİ) haberler bahsettikleri olaylara göre organize etmektedir. Çok sayıda kaynak kullanılmasından kaynaklanan bilgi tekrarlanmasından dolayı YOBİ uygulaması da bazen kendi başına yeterli olamamaktadır. Bu tezde, bir konuyu takip eden haberler üzerinde yenilik bulma (YB) uygulanması incelenmektedir. Bu amaçla ortalama 51 izleyen haber içeren 59 konudan oluşan bir Türkçe YB deney derlemi, BilNov, tarafımızdan hazırlanmıştır. YB için üç metot önermekteyiz; kosinüs benzerliğine dayalı YB yöntemi, dil modellemeye dayalı YB yöntemi ve kapsama katsayısına dayalı YB yöntemi. Ayrıca, literatürde ilk defa kategori temelli sınır değeri öğrenme üzerine de deneyler yapılmaktadır. Ek olarak Türkçe üzerinde YB yöntemleri için doküman vektör uzunlukları ve düzgünleştirme benzeri bazı deneysel parametrelerle ilgili gözlemler sunulmaktadır. Son olarak TREC Yenilik Bulma 2004 deney derlemiyle de deneyler yaptıyoruz. BilNov kullanılarak yapılan deneylerin sonuçlarına göre dil modellemeye dayalı YB yöntemi diğer iki yöntemi belirgin bir şekilde geçmektedir ve ayrıca kategoriye dayalı sınır değeri öğrenme yaklaşımı da genel sınır değeri öğrenmeyle karşılaştırıldığında umut verici sonuçlar vermektedir.

Anahtar sözcükler: Yenilik Bulma, Konu İzleme.

Acknowledgement

I would like to thank to my supervisor, Prof. Dr. Fazlı Can for always being available to me when I needed help. It has been three years since I started working with him and I can't show a single day within these that I didn't enjoy. I learned a lot from him not only about research but also about life.

I also thank to my co-advisor Asst. Prof. Dr. Seyit Koçberber for his comments and helps throughout this study.

I am grateful to my jury members, Asst. Prof. Dr. Pınar Duygulu Şahin, Dr. İlyas Çiçekli and Prof. Dr. Nesim K. Erkip for reading and reviewing this thesis.

I would like to thank to Çağdaş Öcalan and Süleyman Kardaş for showing me the way when I got lost in Bilkent News Portal implementation.

I would like to acknowledge TÜBİTAK for their support under the grant number 108E074. I also thank to Bilkent University Computer Engineering Department for their financial support for both my studies, travels and TREC Novelty Dataset.

I also thank to CS533 - Information Retrieval Systems course students for their helps during creation of BilNov.

I am also grateful to my friends, Abdullah Bülbül, Enver Kayaaslan, Mücahid Kutlu, Tolga Özaslan and Şükrü Torun, with whom I spent two years (more or less) both in the department and the lodgings.

I thank to my family for supporting me with all my decisions and for their endless love. My special thanks to my fiancee, Özlem for supporting me throughout my study and bearing the times of my absence because of my studies and above all, simply for being who she is.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With development of new technologies, amount of digitized information has increased dramatically. In [43], it is claimed that over 90% of information currently produced are generated in a digital format. These contain all types of data such as text, video, audio etc. World Wide Web (WWW) is frequently used for making them accessible.

One of the most commonly shared type of information through WWW is news. Most of the newspapers and news agencies provide news from their web pages. Other than these news providers, news portals also share news by collecting them from the original sources. These news portals gather the news from multiple sources via RSS (Really Simple Syndication) and/or directly crawling. Multi-source news portals provide various advantages such as richness in news content and opportunity to evaluate news from different angles. Additionally, it is practical to follow different news sources from a single web page. Google News (http://news.google.com) can be given as a commercial news portal example. It offers many services such as information retrieval, personalized information filtering, and news clustering. Other research oriented examples are NewsBlaster [28] and NewsInEssence [34] each of which provides clustering and summarization services over the news.

Increasing number of sources and incoming news makes it difficult for news

consumers to find news of their interest in news portals. Different organizational techniques have been introduced in the literature to enable easy browsing in news portals. New event detection and tracking (NEDT), one of these techniques, aims to organize news with respect to the events that they report. An event is defined as a happening that occurs at a specific time and place initiated by the first story reporting the event. As an example, Haiti earthquake is an event. Additionally, a topic is a seminal event. So, a topic is about the developments of a specific earthquake, not all earthquakes. NEDT labels every incoming news to the system as either tracking news of a previous event or the first story of a previously undetected event. Five different problems about NEDT were attacked during Topic Detection and Tracking (TDT) research initiative which was organized between 1997 and 2004 [2], New Event Detection, Topic Tracking, Topic Detection, Story Segmentation and Story Link Detection. Different document similarity calculations were applied by researchers to decide whether a news document is related with any of the previous events or it is the initiator of a new event. Another organizational approach for news portals is Information Filtering (IF). Most of the news portals enable their users to have profiles in which they can save some keywords or documents that reflect their interest area. Some of the algorithms proposed for IF also considers user feedback as input to the system to improve filtering accuracy. IF algorithms basically try to deliver news that are relevant to the users' profiles [8].

1.1 Motivations

Organizational operations enable the news consumers to find news in their interest area easily. However, cluster, event, category or relevance information may not be enough. For example, usage of multiple news sources may cause repetition of the same information within the tracking news of a topic or the delivered news of an IF profile. Sometimes, even a single source may publish several copies of the same news article with small changes. For example, in Google News most of the events have thousands of relevant news from the same or different sources. If all of these news would be served to the news consumers directly, it would be very hard

to follow the developments due to high number of documents on the same topic. All of the provided news may not be interesting for the user because an article may contain no novel information when compared to the previously delivered documents. Documents with novel information should be detected and only they should be served to the user. Allan et al. show novelty detection as a necessary complement to real-world filtering systems because growth of information size raises redundancy as a problem [1]. In Figure 1.1 an example integration of a novelty detection module into a NEDT system is given. After NEDT system gives its tracking decision about the document, d_1, ND system checks whether the document is novel with respect to the previous documents in the topic it is assigned to or not.

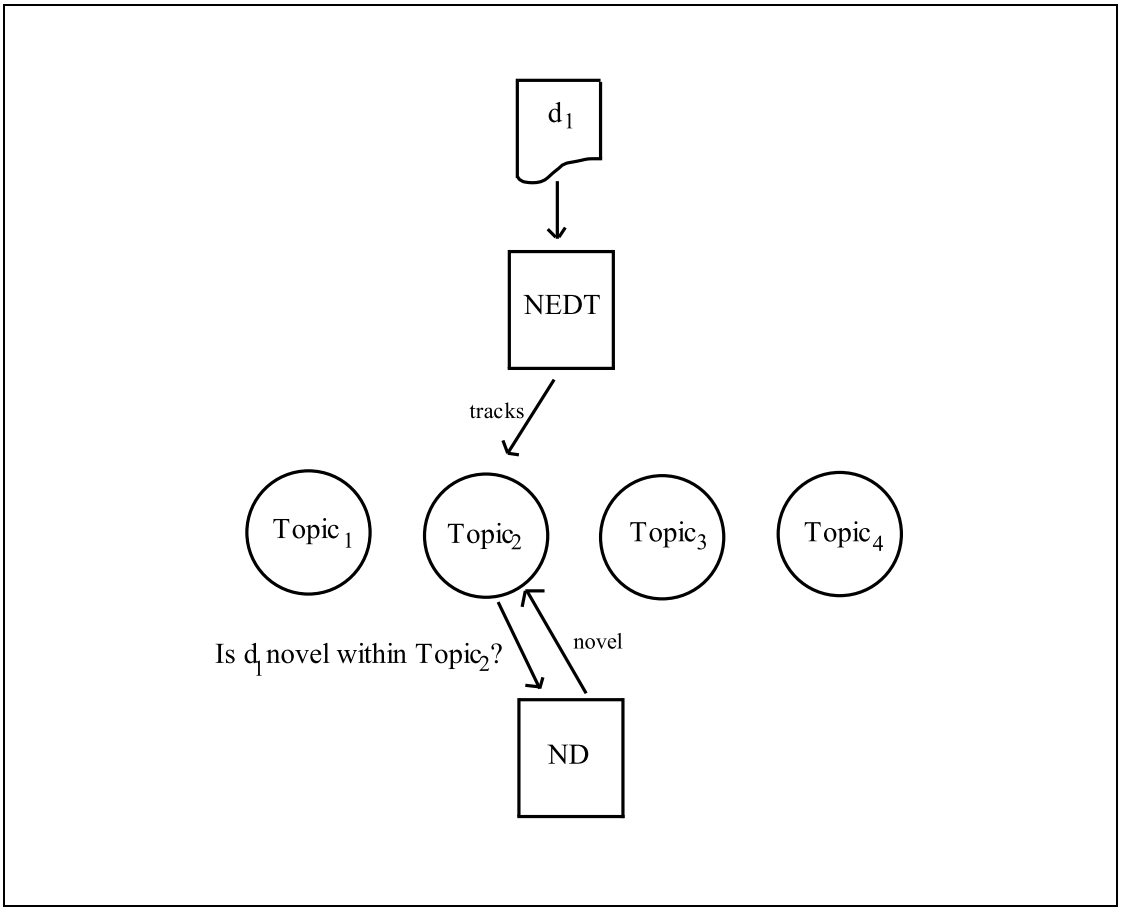

Figure 1.1: Novelty detection module incorporated into a NEDT system .

Novelty detection (ND) may be defined as finding data which contain novel characteristics with respect to some other data. It has been studied in many

domains at different scales with slightly differing problem definitions. In signal processing domain, the task is to identify new or unknown data which has not been encountered during training process [27]. It is also named as outlier detection [17]. In text processing area, ND has been studied in different scales, at event or sentence level.

1. Event level: ND studies at the event level arise from TDT. One of the five tasks of TDT workshop, New Event Detection also called First Story Detection (FSD), was defined as finding the first story that reports an event [2]. In FSD novel information provided in the documents that follow in the timestream is not considered as novelty if they just report developments about the same event. This is why FSD is called event-level ND [26].

2. Sentence level: TREC Novelty Track contains a large body of the work conducted on ND at the sentence level. The workshops were organized between 2002 and 2004. At these workshops, given a set of ranked sentences about a query, the main task was to find relevant and novel sentences. Participants were asked to initially find the set of relevant sentences and then find the set of novel sentences from the set of relevant ones [39]. A sentence is defined as novel if it contains information that was not reported previously in a topic. There were also different tasks which specialize only on relevant sentence retrieval or novel sentence retrieval with differing sizes of training data. There are also other sentence-level ND studies which work on documents such as [48]. In [48], authors define novelty similar to TREC Novelty Tracks. This work is similar to TREC Novelty Tracks except they use documents as the retrieval component and they only work on ND, not relevancy detection.

In this work, we use the novelty definition as in TREC Novelty Tracks. Given the tracking news of a topic, we try to identify documents which contain novel information that was not covered in any of the previous documents. Novelty decision is given for documents. However, systems may make this decision by analyzing the sentences. In Figure 1.2 an illustration of ND problem at this context

is given. Let A, B, C and D represent different information contained by the documents. Red rectangles show the piece of information which causes the document to be regarded as novel. First story is novel by default. Document-1 is novel because it reports information-B which was not reported before. Document-2 is not novel because it contains no novel information. Document-3 report information-C and is novel. Document-4 is not novel and and Document-5 is novel. Document-4 proves another important characteristics of ND problem that it is different than near-duplicate detection [41]. Although both ND and near-duplicate detection aims redundancy elimination, we can see that in the example, Document-4 is neither a near-duplicate of any of the previous documents nor novel. This shows that ND should be handled different than near-duplicate elimination.

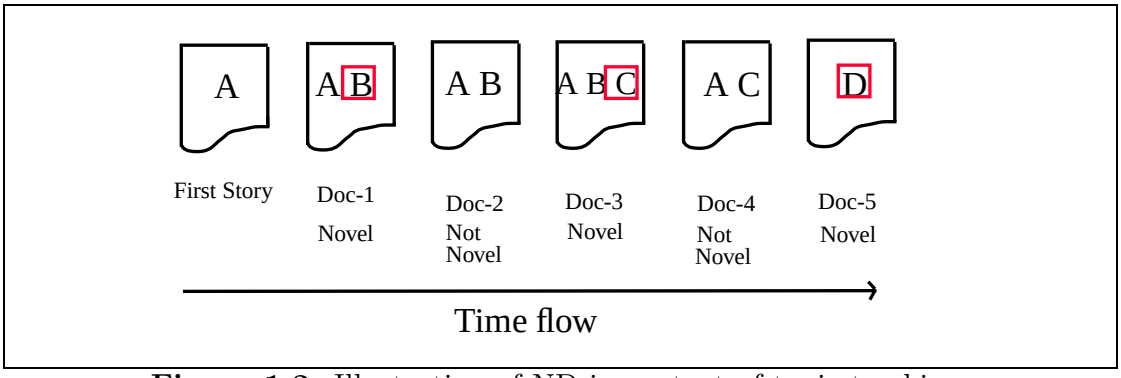

Figure 1.2: Illustration of ND in context of topic tracking.

Dealing with relevancy and novelty at the same time bears a conflicting schema which requires sentences/documents to be similar to the previous ones for relevancy, but also dissimilar for novelty. Since these two tasks are conflicting they should be evaluated separately [48]. In this work we will be working on tracking documents of a topic, so all of the documents are assumed to be relevant to the topic. Even though we work on topic tracking documents, the methods studied in this work can be applied in many other domains such as IF, intelligence applications, patient reports, etc.

1.2 Contributions

In this thesis we:

- Give the details about construction of the first ND test collection in Turkish, BilNov and present some statistics about it;
- Propose usage of three different ND methods; cosine similarity-based ND method, language model-based ND method and cover coefficient-based ND method [11] where first two are adapted from ND literature [5];
- Evaluate performances of the novelty measures using the test collection we constructed and show that language model-based ND methods outperforms the other two methods significantly in terms of statistical tests;
- Experiment on TREC Novelty Tracks' test collections and discuss the differences between the results in Turkish and English;
- Examine the effects of different configurations of a ND system in Turkish such as smoothing methods in language models [20, 46] and document vector lengths in cosine similarity-based method [9];
- Propose usage of category-based threshold learning for ND and compare its results with general threshold learning;

1.3 Overview of the Thesis

The thesis is organized as follows:

- Chapter 2 summarizes the studies on ND by categorizing them as event level, sentence level and other applications.
- Chapter 3 explains ND methods utilized in this study.
- Chapter 4 examines the experimental setup of our study.
- Chapter 5 explains the evaluation measures for ND and presents the results of the proposed ND methods.
- Chapter 6 concludes the discussions and provides some future pointers.

Chapter 2

Related Work

ND studies can be categorized into three classes, event level, sentence level and other applications [26]. In the following sections ND studies at event level, sentence level and in other applications will be summarized respectively.

2.1 ND at Event Level

New event detection problem is introduced in TDT research initiative which was organized between 1997 and 2004 [2]. The problem within the context of a news stream is to find events which were not reported before. There were different tasks introduced in TDT; FSD among these tasks deals with new event detection and is the most similar task to ND at the sentence level.

Different techniques were utilized to attack FSD problem. Clustering was widely used to cluster news which report the same event into the same cluster. This is similar to single pass clustering [42]. 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. This method may be inefficient as the number of clusters increase. Yang et al. proposed sliding-time window concept in which an incoming story is only

compared to the members of a time period [44] which decreases the number of comparisons. They also utilize a time-decay function to lessen the effect of older documents.

Effects of usage of named-entities in TDT systems are also examined. Yang et al. introduce a two-level scheme in which they first classify incoming stories to broader topics like "airplane accidents," "bombings" etc. before performing new event detection [45]. 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 the normal FSD systems which compare incoming stories with all of the document history. Additionally, named-entities are given weights specific to the topics. This is one of the rare studies in which usage of named-entities was significantly better performing. This may be due to the two-level scheme. In [22], although some performance increase is gained by utilizing named-entities, a deeper investigation is suggested. Can et al. report no significant improvement by using named-entities and the authors state that this may be result of the test collection not being conducive to the usage of named-entities [9].

2.2 ND at Sentence Level

Main aims of information retrieval are representation, storage, organization of information and providing easy access to these information. Information retrieval systems, using their underlying organization structure, try to retrieve information that are relevant to a user query [6]. Typically, using a retrieval model, these systems rank the documents in the collection in terms of relevance to the query and provide this ranked list to the user. Increase in the number of documents in the collections brings redundant information problem into consideration. For example, Google's search engine groups very similar pages from a web page and shows only one instance of the page. It provides the users the option to show all of the similar webpages. However, when pages from different sources have the

same information these cannot be detected as similar pages. This redundant information bears the need for a search system which not only detects the relevancy but also novelty.

NIST organized TREC Novelty Tracks between 2002 and 2004 [16, 37, 38]. In these tracks, given a list of documents (split into sentences) that are relevant to a query, there were two defined problems:

- **Relevant Sentence Retrieval**: This problem aims to find sentences which are relevant to the query. Sentence retrieval is considered as different from document retrieval because sentences contain limited amount of text than documents [39]. Since they contain less text, it may be expected that the systems that work on sentences are not reliable. Despite this possible problem, taking sentences as the unit of retrieval enables adjusting sentence-level decisions to different levels of texts such as the aim of these tracks which is a system that helps information retrieval system users to skim through result set of a query by only seeing relevant and novel sentences.

- **Novel Sentence Retrieval**: This problem aims to identify relevant sentences which contain new information with respect to the previous sentences both in the same document and the ones 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 while giving decision for the next sentence. Another important point of novel sentence detection is that, it should be done over relevant sentences. Because new information contained by irrelevant sentences should not be provided to the users. Especially in news this may be encountered very frequently such as sentences which explain some developments related to the event but not directly relevant to the topic or some narrator comments.

Test collections used in TREC Novelty Tracks were consisting of 50 topics each of which contains a query and 25 relevant documents. In TREC 2004, to make the tasks more challenging, some irrelevant documents were also put in the topics. In

Novelty 2002 track, the documents were given in relevance order where in 2003 and 2004 the documents were processed in chronological order which is more appropriate for the nature of ND. Documents were split into sentences by NIST and the annotators were asked to select the set of relevant sentences and within the set of relevant sentences then they selected novel sentences. Performance evaluations were conducted over these ground truth data. As the evaluation measure F-measure was used.

There were 4 different tasks with varying quantities of training data:

1. **Task 1:** Given the set of all documents and the query, find all relevant and novel sentences.
2. **Task 2:** Given the set of relevant sentences, find all novel sentences.
3. **Task 3:** Given the relevant and novel sentences for the first 5 documents, find relevant and novel sentences in the remaining 20 documents.
4. **Task 4:** Given all relevant sentences and novel sentences for the first 5 documents, find novel sentences in the remaining 20 documents.

In the following sections relevant and novel sentence retrieval methods from studies conducted using TREC Novelty Track's test collections will be explained respectively.

2.2.1 Relevant Sentence Retrieval

In TREC Novelty Tracks a variety of relevance measures were utilized for detecting relevant sentences. In most of these methods sentences' similarity to the topic query is used to quantify its relevance. Query expansion methods are also used to make more reliable similarity calculations. In [5, 7] authors expanded the query with the TREC topic definitions and also a proximity-based thesaurus is used for further expansion in the latter one.

Different retrieval models are used for similarity calculations. Vector space model [36] is one of most frequently used models. In this model, texts are represented as N-dimensional vectors where N is the number of unique terms. Value of the dimensions in the vector space model are found by a term weighting function such as $TF - IDF$ [42]. After converting the texts to be compared into vector space model, different similarity measures may be applied. One of these measures, cosine similarity [42] is frequently used in Novelty Tracks [5, 15, 47]. After calculation of similarity, binary relevance decision is given by comparing the similarity with a learned threshold value. For learning the threshold value training data given in different tasks can be used where appropriate, additionally some groups used TREC 2002 and 2003 data for training in 2004 track [5].

In addition, probabilistic models are utilized in relevant sentence retrieval. Language models (LM) are successfully used in information retrieval studies [33]. In this type of retrieval, term statistics of each document are used to estimate a probabilistic unigram model for that document which can be used to find the probability that a word may be generated from the document's model. Maximum likelihood estimator (MLE) for language model proposed by [33] is given in Equation 2.1. In the formula, d stands for the document, θ_d is the model of the document d, t represents the term, tf function gives the frequency of term t in the document, d and $|d|$ represents the length of the document which is the number of tokens in it. As it can be seen in the formula, this estimator gives 0 probability to the terms which are not included in the document. Smoothing methods address this problem by trying to approximate the probability of a term which does not occur in the document. Different smoothing techniques were proposed in the literature [5, 48]. Given LM of two texts, distance of two texts can be calculated via Kullback-Leibler (KL) divergence. This measure is used as a relevance score by negating [5].

$$P(t|\theta_d) = \frac{tf(t, \theta_d)}{|d|} \tag{2.1}$$

Hidden Markov model (HMM), a machine learning approach, is also utilized for relevance detection [14]. An important aspect of this method is that it assumes

fewer independence between documents' relevance. Using the state structure of HMM, relevance of sentence-i may be taken as dependent to sentence-(i-1). HMM requires training for determining state transition, initial state and output probabilities. In tasks where training data was not available, TREC 2003 data was used for training. OKAPI [35] is also utilized to estimate similarity between the query and sentences [47].

2.2.2 Novel Sentence Retrieval

In TREC novelty tracks, a very simple but intuitive method, New Word Count (NWC) [24], was one of the most successful methods. In this method sentences were given a novelty score based on the number of new words that they include. A new word in this context is a word that was not encountered in any of the previous sentences. Like many other methods, this method also needs a threshold value for giving novelty decision.

Similarity measures are also utilized for novelty. The basic idea is to compare a sentence with all of the previous sentences and if the similarity to all of the previous sentences are below a threshold, the sentence is labeled as novel. This idea is adapted from First Story Detection (FSD) in TDT [32]. In [40], cosine similarity is used for similarity calculation. In [15] current sentence is compared with a knowledge repository consisting of all previous sentences instead of all previous sentences one by one. Zhang et al. proposes that since novelty is an asymmetric property, symmetric similarity/distance values should not work well in ND [48]. However, in their study and in most of the studies in the literature cosine similarity which is symmetric was utilized in most of the successful ND methods.

LM are also utilized for novel sentence detection. KL-divergence (see Equation 3.8) is used for measuring the dissimilarity of two LM. Two different ways are followed in [5] which are an aggregate and a non-aggregate method. In aggregate method for giving novelty decision about a sentence KL-divergence between its LM and a LM constructed from all of the previously presumed relevant sentences

is calculated. An aggregate model seems more accurate since a LM constructed from a larger amount of text is more reliable. However, the possible problem about an aggregate model is that redundancy of a sentence may be hidden in an aggregate model. For example a sentence may be regarded as almost a duplicate when compared to a sentence very similar to it, however when compared to a larger set of text which contains the similar sentence, the redundancy of the latter sentence may be hidden. In the non-aggregate method, KL-divergence between models that are built from sentences are calculated. Novelty of a sentence is taken as the minimum KL-divergence value between its LM and the models constructed from the previous sentences. As stated above, the possible problem about non-aggregate method is that sentences may contain very few text and it may be unreliable to construct LM from sentences. Accurate smoothing techniques should be employed to overcome this problem. Different smoothing techniques are used for LM such as Jelinek-Mercer and Dirichlet smoothing. In addition to aggregate and non-aggregate methods, a mixture-model is proposed. Being first introduced by [48], mixture-model 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 address the ND problem in a similar context to question answering [26]. They define 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"), date ("when") are used as question patterns. Then, sentences that have answers to these questions are extracted as novel ones. Problem arises about the opinion topics whose queries do not include such patterns. Different patterns such as "states that" are proposed for opinion topics. Additionally, a detailed information pattern analysis of sentences in TREC novelty data is given in the paper.

2.3 Other applications

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

In Zhang et al. an adaptive filtering system is extended for redundancy elimination [48]. Documents to be delivered for a filtering profile is processed by redundancy elimination tool and documents which are redundant given the previously delivered documents are eliminated. Experiments on different measures are conducted in this study. Authors claim that since novelty is an asymmetric measure (when documents are reordered, a novel document may be not novel), symmetric measures should not be performing well. However, one of the best performing methods were a cosine similarity-based method adapted from FSD and the other one was a mixture of LM.

ND at sentence level has many similarities with summarization studies. In both only the necessary sentences should be delivered to the user. In summarization there is also 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, in summarization, because of compression concerns, only the latter sentence may be contained in the summary. A subtopic of summarization area, temporal summarization, aims to generate summary of a news stream timely, considering the previous summaries and providing only the updates from the previously delivered summary. Allan et al. define usefulness (which may be understood as similar to relevancy) and novelty of sentences and tries to extract novel and useful sentences [3]. Language modeling is used with a very simple smoothing technique. Additionally, update summarization is a similar problem which was piloted in Document Understanding Conference 2007 and continued in Text Analysis Conference 2008 and 2009. The aim in update summarization was to generate a summary of a set of documents under the assumption that another set of documents are already read by the user.

Temporal text mining deals with analyzing temporal patterns in text. In [29], evolutionary theme patterns are discovered. As an example, in a text stream related to Asian tsunami disaster, the aimed themes are "immediate reports of the event," "statistics of death," "aids from the world" etc. Also, a theme evolution graph is extracted in which transitions between themes are shown. LM are also utilized in this study. Parameters of the probabilistic models are estimated by Expectation Maximization algorithm [30].

Chapter 3

ND Methods

In this section our proposed ND methods are explained. Prior to application of these methods, some pre-processing methods are applied on the texts which are explained in Section 3.1. Following the Pre-processing section, we explain category-based threshold learning approach in Section 3.2. In Section 3.3, random baseline, cosine similarity-based ND method, language model-based ND method and cover coefficient-based ND methods are explained respectively.

3.1 Pre-processing

Natural language products cannot generally be used by computer applications directly, some pre-processing should be applied to the text. There are generally three steps of preprocessing:

- Tokenization
- Stopword Elimination
- 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 spaces and punctuation marks.

3.1.1 Stopword Elimination

In information retrieval studies words are generally given some importance with respect to their frequency in the text. Stopwords may affect performance of these studies since they generally occur very frequently in texts. Stopword elimination is applied to texts before processing in order to overcome this effect. Since these words do not distinguish sentences/documents from each other, elimination of them is expected to increase system performance.

In Turkish information retrieval effects of stopword elimination is examined [10]. Authors utilize three stopword lists and report no significant difference between effectiveness of different configurations. As a more similar study to ND, Can et al. also show that using a stopword list significantly increases the effectiveness in new event detection [9]. However, there is no significant difference between effectiveness of the system with longest stopword list and the system with a shorter list.

In this work we utilize the longest stopword list which contains 217 words taken from [21]. This is a manually extended version of a shorter stopword list [9].

3.1.2 Stemming

In natural languages, prefixes and suffixes are used to either derive words with different meanings or inflect the existing words. Different stemming algorithms are used to find the stems of the words so that word comparisons may be more reliable. In this work we utilize a stemming heuristic called Fixed Prefix Stemming.

Turkish is an agglutinative language in which suffixes are used to obtain different words (a more detailed characteristics of Turkish is given in [25]). In fixed prefix stemming, words' first N characters are used as the word stem. For example, for word *ekmekçi* (bread seller), first-five(**F5**) stem of the word is *ekmek* (bread). Turkish's agglutinative property makes fixed prefix stemming an appropriate approach. Can et al. [10] showed that F5 stemming gives the best performance in Turkish information retrieval. Additionally, in new event detection it is shown that systems using F6 is one of the best performing ones [9]. In this study we will utilize F6 stemming with the help of observations done in [9].

3.2 Category-based Threshold Learning

We utilize cross validation for reporting our system performance since all of our methods have some parameters and these should be learned. In this study, motivated from [45] we also try category-based threshold learning and compare results of general threshold learning with category-based threshold learning. In [45], the authors study running FSD on a local history based on a category instead of all of the previous documents. Our motivation here is that each topic has a different type like sports news, accident news etc. and each of these categories have different novelty structures. For example, intuitively, one would expect to see more rapid but small developments in an accident topic where in a topic related to politics it may take days for the topic to become mature. So, we hypothesize that while learning a threshold for a topic, if we use only topics from the same category with the topic, we can increase the system performance. In our test collection there are 13 different categories such as accidents, financial etc. We experiment with category-based threshold learning using these categories. We report the results of category-based threshold learning Section in 5.2.1.5.

3.3 ND Methods

3.3.1 Baseline - Random ND

Systems which give their decisions randomly are widely used as a baseline in many problem areas [18]. With comparison with this baseline, a method can be proven to work better than a random baseline and its decisions are justified as different than just random decisions.

In ND context, the random baseline method is straightforward, without examining the contents of a document, it gives novel/not novel decisions with a probability of 0.5. In order to evaluate random baseline, expected performance of the system 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 3.1.

Documents in Topic K	$\boxed{1}$	2	3	$\boxed{4}$	5	.....	m
Probability of Being Labeled as Novel	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	.....	$\frac{1}{2}$
Contribution to Recall	$(\frac{1}{2} * 1) +$	$0 +$	$0 +$	$(\frac{1}{2} * 1) +$	$0 +$	..... +	$0 = \frac{a}{2}$

Figure 3.1: Calculation of expected performance of random baseline.

Let K be a topic with m documents as in Figure 3.1 and a be the number of novel documents in these m documents. First row of the figure shows the documents, documents surrounded with a red square are novel documents. The second row shows the probabilities of each document being labeled as novel. As we stated, this probability is 0.5 for all document in random baseline. Third row shows the contribution of each document to recall if they 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 1 contribution to the measures and they can be involved in the set with 0.5 probability, so in the expected case sum will be $\frac{a}{2}$. So, we can derive recall as R = $\frac{\frac{a}{2}}{a} = \frac{1}{2}$. However, for precision the contribution of a document is not only to the numerator part of the formula, also the denominator part of precision formula increases (recall calculation can be done easily as we did since denominator part of recall is constant, A). So, we give a general formula for precision calculation in Equation 3.1 for a topic with m documents and a novel documents where $a > 1$. 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 j documents can be chosen from m-a not novel documents. Precision at this case is $\frac{i}{i+j}$ which is the ratio of novel documents in the set of returned documents. The denominator 2^m is the number of total cases (it might also be taken as $2^m - 1$ since the case where no documents are returned, precision is not defined but we neglect this).

$$Precision = \frac{\sum_{i=1}^{a} \sum_{j=0}^{m-a} \binom{a}{i}.\binom{m-a}{j}.\frac{i}{i+j}}{2^m} \tag{3.1}$$

Results of random baseline will be given in Section 5.2.1.1.

3.3.2 Cosine Similarity-based ND

In many text-based studies 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. Cosine similarity is one of the most frequently used similarity measures in information retrieval. Its geometrical interpretation is that it is equal to the cosine of the angle between two vectors. In text similarity calculation texts to be compared are initially converted into vector-space model [36]. In this model, every unique term is represented by a dimension in the vectors and the value 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 stands for term frequency and IDF stands for inverse

document frequency. Calculation of *TF-IDF* value of a term in a document is given in Equation 3.2. In the equation $tf(t,d)$ is the frequency of term t in document d. Second part of the multiplication is IDF part in which m represents the number of the documents in the collection and m_w is the number of documents which contain term t. 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 the initial results obtained with $TF - IDF$ function. Cosine similarity tends to give good results even just with raw term frequencies. Similar observations were reported in [4].

$$TF - IDF(t,d) = (tf(t,d).log(\frac{m}{m_w})) \tag{3.2}$$

Formula of cosine similarity is given in 3.3. In the numerator dot product of the vectors, w_i and w_j are calculated by summing the multiplication of corresponding dimensions. denominator is a normalization factor which consists of multiplication of lengths of both of the vectors. N is the number of dimensions in both of the vectors.

$$CosSim(d1,d2) = \frac{\sum_{k=1}^{N} w_{ik}.w_{jk}}{\sqrt{\sum_{k=1}^{N} w_{ik}^2 . \sum_{k=1}^{N} w_{jk}^2}} \tag{3.3}$$

Our cosine similarity-based method is adapted from FSD. In this algorithm we identify a novel document as a document which is dissimilar to all of the previous documents to an extent. Comparisons should be made with all of the previous documents because high similarity to even a single document may make a document not novel. The algorithm can be seen in Algorithm 3.1. Document arriving at time t, d_t is compared to all of the previous documents and if its similarity to any of the previous documents is greater than threshold, θ, the document is labeled as not novel. Otherwise, the document is labeled as novel.

Algorithm 3.1 Cosine Similarity-based ND Algorithm

```
1: d_t is the document arriving at time t
2: θ is the novelty threshold
3: for Every previous document d do
4:    if CosSim(d_t, d) ≥ θ then
5:       d_t is not novel
6:       RETURN
7:    end if
8: end for
9: d_t is novel
```

3.3.2.1 Reduction of Document Vector Length

Document vector length of a document is the number of unique terms that the document contains. In other words it is the number of non-zero valued dimensions in the document vector. In cosine similarity normally all terms of texts (dimensions of vector) are used for the calculation. Using all dimensions does not necessarily make similarity calculations more reliable since some terms with smaller frequency may not make contribution to the similarity between documents. For example, even after stopwords are eliminated, some topic specific stopwords may exist in the documents and these may cause the documents to be assumed as more similar to each other than they actually are. Even though Allan et al. state that cosine similarity tends to perform better at full dimensionality [4], document vector length is an important feature which should be examined. Effects of document vector length were studied in new event and detection [9]. We evaluate effects of using different document vector lengths (highest valued dimensions) in cosine similarity calculation in ND in Section 5.2.1.2.

3.3.3 Language Model-based ND

Probabilistic models have been incorporated in information retrieval for over four decades [46]. These models try to estimate the probability that a document is relevant to the user query [33]. Ponte and Croft [33] introduced a new and simple probabilistic approach based on language modeling. This new model unlike its

predecessors does not have any prior assumptions on documents such as coming from 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 [33] is given in Equation 3.4. In the formula, $tf(t, d)$ is the term frequency function which gives the number of occurrences of t in document d and $|d|$ is the length of 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 0 with MLE. This is a very strict decision and generally does not reflect the true probability of the term.

$$P_{MLE}(t|\theta_d) = \frac{tf(t, d)}{|d|} \tag{3.4}$$

Smoothing methods aim to empower MLE of the probabilities so that unseen terms in the documents are not assigned 0 probability. Especially, when estimating a model with limited amount of text, smoothing has a significant contribution in model's accuracy [46]. Allan et al. apply smoothing in a simple way by adding 0.01 to numerator of P_{MLE} and multiplying denominator by 1.01 [3]. This approach helps to overcome problems caused by unseen terms, however it does not offer a good estimate of the probability. Zhai and Lafferty [46] examine different smoothing methods for information retrieval. In this study, we will experiment with two different smoothing methods which are Bayesian Smoothing Using Dirichlet Priors and Shrinkage Smoothing [5, 46].

3.3.3.1 Smoothing Methods

1. **Bayesian Smoothing Using Dirichlet Priors:** This smoothing method which is also called Dirichlet Smoothing is similar to Jelinek-Mercer smoothing [20] because it also uses a linear interpolation of MLE model with another model. Model obtained by Dirichlet smoothing is given in Equation 3.5. In the equation, $tf(t, d)$ is the count of occurrences of t in document d, $P_{MLE}(t|\theta_C)$ is a MLE model constructed from a collection of documents

C to smooth the probability of the document model and μ is 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. So, basically a term's probability of generation from the document model will depend on its probability of occurrence in the previous documents. Dirichlet smoothing takes language models as multinomial distributions whose conjugate prior is a Dirichlet distribution [46] and parameters of Dirichlet distribution are taken as $\mu p(t_1|\theta), \mu p(t_2|\theta), \mu p(t_3|\theta), \mu p(t_4|\theta), ..., \mu p(t_n|\theta)$. Another property of this smoothing as can be seen by the weights of the components of interpolation, it tends to smooth shorter documents more than the longer documents [5]. In this smoothing model, μ is obtained with training.

$$P(t|\theta_d) = \frac{|d|}{|d| + \mu} P_{MLE}(t|\theta_d) + \frac{\mu}{|d| + \mu} P_{MLE}(t|\theta_C) \tag{3.5}$$

2. **Shrinkage Smoothing:** This method assumes that each document is generated by contribution of three language models, a document model, a topic model and a background model, in our case a Turkish model. Equation 3.6 illustrates the shrinkage smoothing. In this equation, $P_{MLE}(t|\theta_T)$ is the MLE model generated for the topic of document d and $P_{MLE}(t|\theta_{TU})$ is the MLE model generated for Turkish. Interpolation weights for the corresponding LM are shown as $\lambda_d, \lambda_T, \lambda_{TU}$ where $\lambda_D + \lambda_T + \lambda_{TU} = 1$. In our experiments, $P_{MLE}(t|\theta_T)$ is generated by the topic description which is expanded by the first story of the topic. Allan et al. also used TREC topic descriptions for topic models [3]. Turkish model, $P_{MLE}(t|\theta_{TU})$, is generated by using a reference collection, Milliyet Collection [10], which contains about 325,000 documents which are news from the Milliyet newspaper between the years 2001 and 2004. This corpus was utilized in other studies for IR experiments [10] and again as a reference corpus for calculation of IDF statistics [9].

$$P(t|\theta_d) = \lambda_d P_{MLE}(t|\theta_d) + \lambda_T P_{MLE}(t|\theta_T) + \lambda_{TU} P_{MLE}(t|\theta_{TU}) \tag{3.6}$$

3.3.3.2 Adaptation of Language Models to ND

Language models have been used as novelty measures previously in different studies. In [3], occurrence of words in sentences are assumed independent and probability of a sentence s being generated by a model θ is calculated as in Equation 3.7 where t represents terms and s represents sentences. Later these values are directly used as novelty scores. This method seems to depend heavily on quality of smoothing since one unrealistic (small) probability can make the result unreliable because of the multiplications.

$$P(s|\theta) = \Pi_{t \in s} P(t|\theta)^d \tag{3.7}$$

Kullback-Leibler (KL) divergence is another measure used for utilizing language models in ND. KL divergence is used to find distance between two probabilistic distributions. Calculation of KL divergence between two language models, θ_1 and θ_2 are given in Equation 3.8. 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 have the same values. This property makes it a more appropriate measure for ND.

$$KL(\theta_1, \theta_2) = \sum_w P(t|\theta_1).log\frac{P(t|\theta_1)}{P(t|\theta_2)} \tag{3.8}$$

In this study, we also utilize KL divergence as the novelty measure for language model-based ND. In previous ND studies two different ways were followed which are aggregate and non-aggregate methods [5, 48]. In aggregate method, while giving novelty decision of a sentence, all of the presumed relevant sentences were used to form an aggregate model and KL divergence between model of the sentence and this aggregate model is calculated as the sentence's novelty score. While this model seems more accurate because of the larger amount of text, a possible problem is that redundancy of a sentence may be hidden in an aggregate model. For example, a sentence may be regarded as almost a duplicate when compared to a sentence very similar to it, however when compared to a larger

set of text which contains the similar sentence, the redundancy of the latter sentence may be hidden. This problem is also valid for our case, so we utilize the non-aggregate method. In the non-aggregate method, we calculate KL divergence between models of every document separately. Novelty of a document is taken as its minimum KL divergence value with the previous documents. Details of the algorithm are given in Algorithm 3.2. For an incoming document, d_t, we calculate KL divergence between every previous document, if KL divergence between d_t and any of the previous documents is less than the threshold, Θ, d_t is labeled as not novel. This comparison has similar intuitions as cosine similarity-based method except KL divergence is a distance measure.

Algorithm 3.2 Language Model-based ND Algorithm

```
d_t is the document arriving at time t
Θ is the novelty threshold
for Every previous document d do
    if KL(θ_{d_t}, θ_d) ≤ Θ then
        d_t is not novel
        RETURN
    end if
end for
d_t is novel
```

3.3.4 Cover Coefficient-based ND

Cover coefficient (CC) is a concept to quantize the extent to which a document is covered by another document [11]. CC is calculated as in Equation 3.9.

$$c_{ij} = \sum_{k=1}^{n}[\alpha_i.d_{ik}].[\beta_k.d_{jk}] \quad where \quad \alpha_i = [\sum_{l=1}^{n} d_{il}]^{-1} \quad \beta_k = [\sum_{l=1}^{m} d_{lk}]^{-1} \tag{3.9}$$

In the formula, n and m, respectively, represent the number of terms and documents in the document-term matrix, D, of a set of documents. Values d represent D matrix entries, i.e. d_{ik} is the number of occurrences of term-k in

document-i where $1 \leq i \leq m$ and $1 \leq k \leq n$. Reciprocals of i-th row sum and k-th column sum of D matrix are represented as α_i and β_k respectively.

Coverage of document-i by document-j, $c_{ij}(1 \leq i, j \leq m)$, 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 construction of C matrix is given in Figure 3.2 which is adapted from [9]. The leftmost part shows an example document-term matrix which consists of 5 documents $(d_1, d_2, d_3, d_4, d_5)$ and 4 terms (t_1, t_2, t_3, t_4). As stated in [11], all documents should at least have one non-zero entry in D matrix meaning that they should contain at least one term and each term should at least be contained by one document. D matrix contains binary values in this example but it may also have the frequencies of the terms in the corresponding documents instead of binary values. In the middle part of Figure 3.2, an example of 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 probability of both terms are 0.5 (obtained by α_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 β_4). This stage is handled by the second part of the formula. The last part of the figure shows the constructed C matrix, a mxm matrix, from the D matrix which contains the c_{ij} values.

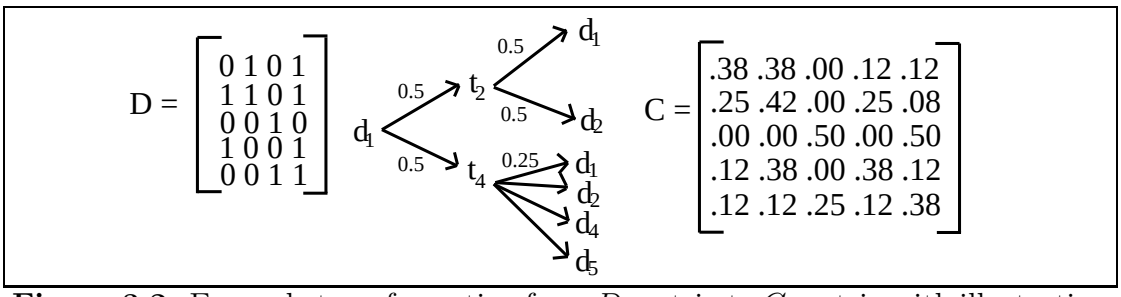

Figure 3.2: Example transformation from D matrix to C matrix with illustration of the term selection probabilities.

3.3.4.1 Motivation for Usage of CC as a Novelty Measure

CC values are probabilities and show the characteristics of probabilistic observations. All c_{ij} values may have values between 0 and 1. If two documents contain no common terms, coverage of one by the other one is 0. Likewise, if only two documents are considered and they are duplicates, their coverage values are 1. Also again if only two documents are considered, and one is a subset of the other one, its coverage by the superset is also 1. Row sum of C matrix is equal to 1 which shows that sum of probabilities of a document covered by the other documents are equal to 1. A document's coverage of itself is called decoupling coefficient and showed by c_{ii} value for $1 \leq i \leq m$. If a document contains terms which only exists in itself, decoupling coefficient of the document is 1 and coverage value by all other documents are 0.

CC value is an asymmetric measure which can easily be shown by an example of two documents in which one of the documents contain the other one. Coverage of the smaller document by the superset is 1 where coverage of superset by the subset is a number smaller than 1. This asymmetric property makes CC concept useful as a novelty measure because same situation exists in ND also. Consider two documents d_1 and d_2 as in Figure 3.3 which may be regarded as tracking documents in a topic. Information contained by the documents are 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. So, d_1 is novel. At time t_2, d_2 arrives and it contains information A and B. Information-B was not reported before t_2 so this document is also labeled as novel. To observe the asymmetry property, we swap the order of 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 was already given in d_2 before. This property may not be handled well by symmetric similarity measures such as cosine similarity since similarity between d_1 and d_2 is calculated regardless of their arrival times. In CC, coverage of d_1 by d_2 will be expected to be larger than the coverage of d_2 by d_1 in this specific case which satisfies the ND property.

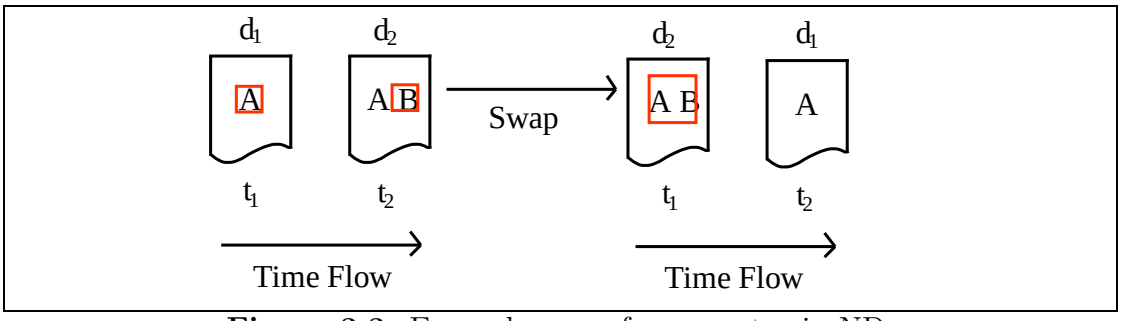

Figure 3.3: Example case of asymmetry in ND.

3.3.4.2 Adaptation of CC to ND

CC may be regarded as an asymmetric similarity measure as explained in Section 3.3.4.1. Since for a document to be novel, we will look for the condition that its similarity to all of the previous documents is below a threshold value. Here, comparisons with all previous documents is important because a document may be dissimilar to almost all of the documents but if it is very similar to even one document, it cannot be labeled as novel. Basic algorithm can be seen in Algorithm 3.3. As it can be seen in lines 3,4 and 5, if d_t is covered by any of the previous documents, d, to an extent, it is considered directly as not novel and similarity calculations are stopped. If all of these comparisons are successful in terms of comparison to the threshold θ, d_t is labeled as novel.

Algorithm 3.3 Cover coefficient-based ND algorithm.

1: d_t is the document arriving at time t
2: θ is the novelty threshold
3: **for** *Every previous document* d **do**
4: **if** $c_{d_t,d} \geq \theta$ **then**
5: d_t is not novel
6: RETURN
7: **end if**
8: **end for**
9: d_t is novel

Threshold θ is learned by cross validation in our experiments. Details of training process are explained in Section 4.3.

Chapter 4

Experimental Environment

In this section we will explain our experimental setup. Details about construction of Turkish ND test collection will be explained in Section 4.1. Later, we will explain TREC Novelty Track 2003-2004 test collections. Finally, we will give some information about our training approach in Section 4.3.

4.1 BilNov - Turkish ND test collection

There are no previous ND studies in Turkish and this poses the problem that there is no standard test collection for objective performance comparison between the methods that will be developed for Turkish. In this section, we report the construction details of the first Turkish ND test collection, BilNov. To the best of our knowledge, this test collection is one of the first ND test collections constructed on tracking news of topics.

BilNov is based on a TDT collection, BilCol2005 [9]. In TDT context a topic is about a development which is triggered with a first story and is followed by the trackers of the first story which are other news related to the topic. A list of example topics are given in Figure 4.1. First row contains a topic about Turkey's first septuplets. First story of the topic has 17.02.2005 as timestamp and 56 news

Table 4.1: Topic examples.

Title	Category	Time Span	# of Trackings
Turkey's First Septuplets	Celebrity/Human Interest	17.02.2005 14.12.2005	56
New Turkish Criminal Code	New Laws	01.06.2005 10.12.2005	53
Trial of Saddam Hussein	Legal/Criminal Cases	10.12.2005 28.11.2005	80

documents related with this topic track it. Last of these tracking documents has the timestamp 14.12.2005. Judgments of first stories and the tracking documents are made by human annotators and the details of annotation process are given in [9].

4.1.1 Selection of Topics Used in the Collection

BilCol2005 collection consists of 80 topics with an average of 72 tracking news. Although, average number of trackings is 72, there are both topics with few trackings and with a lot of trackings such as 245 documents. Our initial experience on annotation process showed that topics with large number of tracking documents are very hard to annotate because with each document, size of information that the annotator should remember increases and also as the amount of time spent during annotation increases, possibility of making mistakes also increases. Other than very long topics, small topics would not be appropriate for ND task because they are not challenging enough to be used in performance evaluation. Because of these reasons, we chose 59 topics from BilCol2005 which contains more than or equal to 15 tracking documents. We only use the first 80 documents of the topics which contain more than 80 documents for the topic length considerations. Figure 4.1 illustrates the distribution of topic lengths in BilNov. As the figure shows, there are plenty of topics from varying lengths which may help the researchers during evaluation of their methods in terms of topic lenghts.

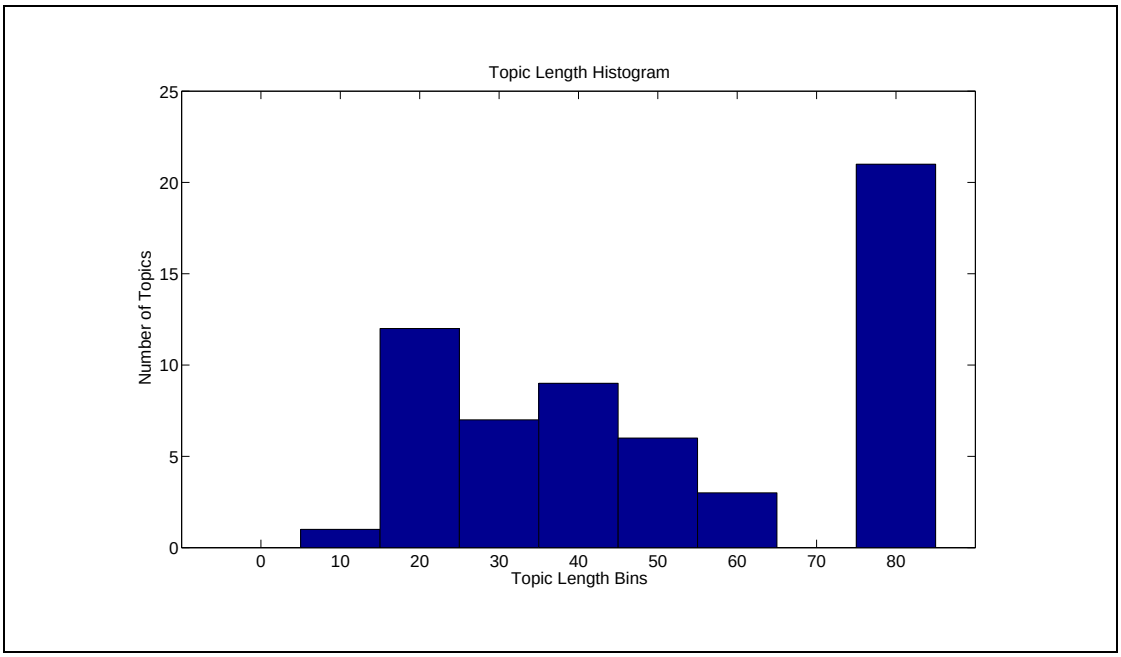

Figure 4.1: Histogram illustrating the distribution of topic lengths.

4.1.2 Annotation Process

Documents are annotated by human annotators within the time sequence (each document has a timestamp). An annotator starts reading from the first story of a topic and then reads all of the documents in the topic in time sequence. After reading each document (except the first story), annotator gives the decision whether the document is novel or not with respect to the previous documents. As the annotation software, we built a component for a previous annotation system, E-Tracker [31]. A screenshot of annotation interface can be seen in Figure 4.2. We worked with 38 different annotators each of which are assigned different number of topics but we tried to keep the total number of documents annotated by an annotator same.

We also asked the annotators to enter the time they spent per topic. Average time spent per a topic is 59 minutes which shows the hardness of the job when performed by a human. Statistics about the test collection is given in Table A.1.

Figure 4.2: Screenshot showing the annotation screen.

4.1.3 Construction of Ground Truth Data

In the literature, generally more than one annotators are used on the same subject to see the effect of having different people assessing the same subject. Although these different judgments may be used separately to observe two different point of views, generally a single ground truth data is generated by using judgments of different annotators.

In our study, each topic is annotated by two annotators. Majority voting would not work obviously in this case since no majority can be obtained when there is a disagreement with two decisions. In some studies, different annotators are asked to work together to decide on one of the decisions. This process is also very time demanding. In their work, Zhang et al. [48] instructs the annotators to give novelty decisions at three level; absolutely novel, somewhat novel and not novel. Later, they conduct 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. We follow a similar approach to Zhang et al. by combining decision of the annotators. If we neglect annotator mistakes, the disagreement between the decisions is probably caused by different interpretations of novelty. So, if we combine decisions of annotators in two different setups, we would be

able interpret novelty in different dimensions. These two configurations are as follows:

- **Optimistic ground truth:** In this ground truth data, when two annotators are in disagreement, we choose decision which is more optimistic about novelty of the document. In other terms, if one of the decisions is "novel", the optimistic ground truth label is also novel. This is similar to logic function, OR, if we consider novelty as 1, if any of the decisions is a 1, the optimistic ground truth is also 1.
- **Pessimistic ground truth:** In this ground truth data, contrary to the previous one, 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 (not novel).

4.1.4 Quality Control of Experimental Collection

Construction of experimental collections requires dealing with lots of data and it is very hard do examine these one by one to evaluate their appropriateness for the task that the collection is built for. During and after the construction, generally some quality control techniques are applied to both the data and the judgments. With the help of these techniques an error about the collection may be corrected or some topics, document which have undesired properties may be eliminated.

In the following three sections, we will explain some analysis of data we performed for quality check.

4.1.4.1 Analysis of Topic Lengths

Lengths of topics are important for a ND collection. A test collection built from very short topics could not effectively be used in performance measure since even a random method can perform well because of the few number of documents. Additionally, choosing topics at same length (all long or all short) could hide

some performance degradation of methods towards some kind of topics. We gave the distribution of lengths of topics included in BilNov in Figure 4.1. As it can be seen, there are topics of different lengths and also there are not any very short topics.

4.1.4.2 Analysis of Novelty Ratios

Novelty ratio is defined as the ratio of the labeled documents which are novel. As a quality feature, it gives us information about the structure of the test collection. A test collection with a higher novelty ratio can be considered as a less challenging test collection since after some ratio it may be more meaningful to label all documents as novel (equivalent to not performing ND). While calculating novelty ratios, since there are two judgments, we took average of them. Distribution of novelty ratios is given in Figure 4.3.

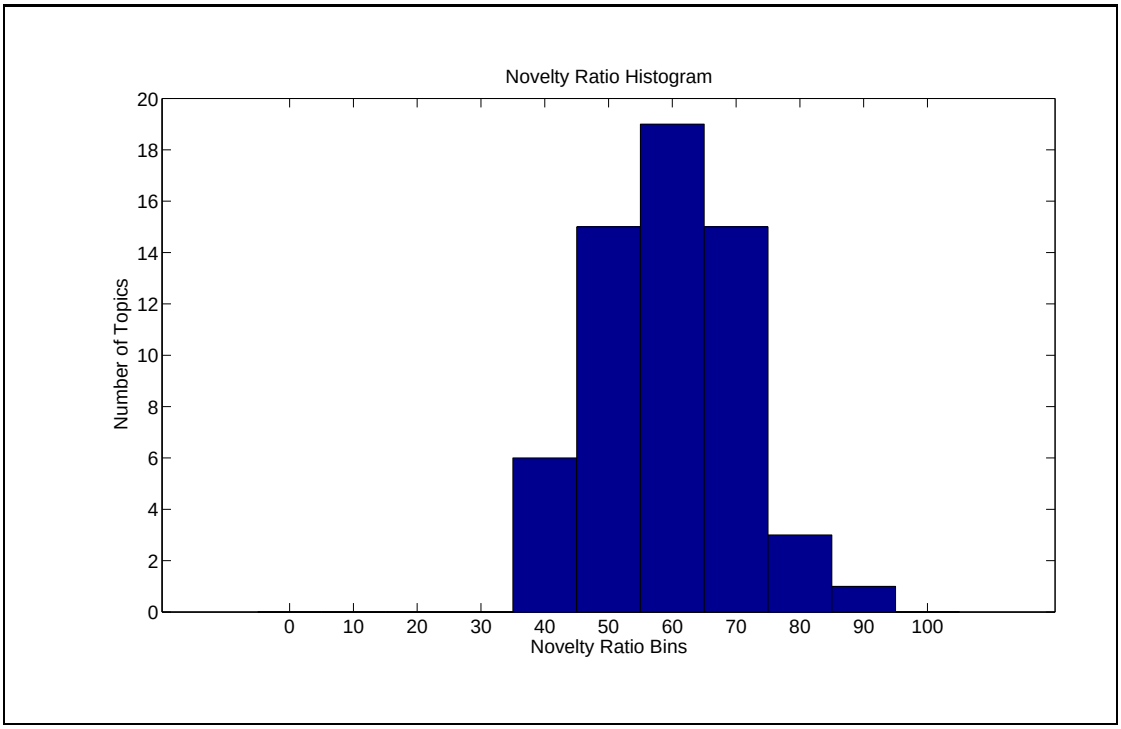

Figure 4.3: Distribution of novelty ratios.

4.1.4.3 Inter-Annotator Agreement

Reliability of the ground truth data constructed from the decisions of different annotators depends on the agreement between the annotators. Kappa coefficient is widely used for measuring inter-annotator agreement [12]. Kappa's superiority to different measures is that it also checks for agreement by chance. Agreement of the annotators is corrected by the expected value of agreement between annotators which is again calculated by using the probabilities of cases obtained from annotator decisions. Formula of Kappa is given in Equation 4.1. In the formula Agr stands for the observed agreement between the annotators. $E(Agr)$ is the expected agreement which calculated by 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 role as a normalization factor.

$$\kappa = \frac{Agr - E(Agr)}{1 - E(Agr)} \tag{4.1}$$

An example case is given in Table 4.2. Rows represent the decisions of annotator A and columns represent annotator B. Expected agreement between the annotator is calculated by 0.75 * 0.4 + 0.25 * 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 both labeled as novel or not novel. So Kappa value is, $\frac{(0.35+0.20)-0.45}{1-0.45} \simeq 0.18$. Kappa coefficient takes values less than or equal to 0 for cases where there is not a agreement more than the expected case. In case of perfect agreement, it takes the value 1.

In our judgments, the average Kappa coefficient is 0.63. This value stands for a substantial agreement according to intervals given by Landis and Koch [23]. Additionally, we performed the statistical test proposed by Conrad and Schriber [13]. In this test, we showed that our Kappa value is significantly different than 0 with p = 0.002. This shows that our agreements are significantly larger than the expected cases.

Table 4.2: Example case for Kappa calculation between annotators A and B.

Annotators' Judgments		B		
		Novel	Not Novel	Total
A	Novel	35	5	40
	Not Novel	40	20	60
	Total	75	25	100

4.2 TREC Novelty Track 2003-2004 Test Collections

In order to evaluate Cover Coefficient-based ND method in other languages, we utilize TREC 2003 and 2004 test collections. A brief introduction to TREC Novelty Tracks were given in Chapter 2. NIST organized Novelty Tracks between 2002 and 2004 years. The aim of these tracks were to go beyond classical relevance based search engines and making search engines able to both determine relevant and novel information. There were four task defined on a ranked list of documents with respect to a query:

1. **Task 1:** Given the set of all documents and the query, find all relevant and novel sentences.
2. **Task 2:** Given the set of relevant sentences, find all novel sentences.
3. **Task 3:** Given the relevant and novel sentences for the first 5 documents, find relevant and novel sentences in the remaining 20 documents.
4. **Task 4:** Given all relevant sentences and novel sentences for the first 5 documents, find novel sentences in the remaining 20 documents.

Since relevancy detection is out of scope of this study, we focus on Task 2 which gives the set of relevant sentences and asks to find all novel sentences. We utilize 2003 data for training and 2004 data for testing. We did not use 2002 data since it was a little problematic. During 2002 track, very few sentences were

chosen relevant and as a result of this, almost all of the relevant sentences were chosen as novel by the annotators. Additionally, since it was the pilot year, the definition of the problem was not clear enough, i.e. the documents were processed in relevance score order, not chronologically.

More detailed information on TREC Novelty Tracks can be found in [16, 26, 37, 38].

4.3 Training

All of our methods has some parameters and these should be picked properly to evaluate systems' performance. In Turkish ND experiments, we follow two different ways:

1. **30-fold cross validation:** In our general threshold learning experiments we report our results as the average of fold performances with 30 folds. In this approach our test collection is split into 30 parts and each of these parts is used once as for testing where the rest are used for learning the threshold values.

2. **Leave-one-out cross validation:** In our category-based threshold learning experiments, since categories may contain few topics, we apply leave-one-out cross validation.

Chapter 5

Evaluation Measures & Results

In this chapter we first explain the evaluation measures used in this study in Section 5.1. In Section 5.2 we report the evaluation results of our methods and discuss them.

5.1 Evaluation Measures

In TREC Novelty Tracks F-measure was used as the evaluation criterion [16, 37, 38]. Soboroff and Harman discussed the possible problems which may be encountered if precision and recall are directly used for evaluation [39]. To overcome these problems, F-measure was used. If we want to give equal weights to precision and recall, F-measure can be calculated like the following where P stands for precision and R stands for recall.

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

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.

Table 5.1: 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

Zhang et al. [48] in their redundancy elimination study both used standard IR measures, precision and recall and also used mistake as a measure which is generally used classification problems.

In this study we will also use F-measure as in TREC Novelty Tracks. All results we report throughout this study are macro-averaged, they are calculated for each topic and then averaged over all topics.

5.2 Evaluation Results

5.2.1 Turkish ND Results

5.2.1.1 Random Baseline Results

In this section, we present the results of the random baseline system. Results can be seen in Table 5.1. F-measure values are not directly calculated from average of the precision and recall values instead they are calculated for each topic and then averaged. We can see that the random baseline performs as expected. As we stated before, in a challenging test collection random systems should not be able to perform well. In pessimistic test collection, performance of random degrades since disagreement values are taken as not novel, there appears to be less novel documents. In the following sections, we will compare results of the proposed methods with the random baseline to see how well they perform.

5.2.1.2 Cosine Similarity-based ND Results

In our cosine similarity method, we experimented with different document vector lengths. Results of these experiments for optimistic and pessimistic test collections are in Table 5.2 and Table 5.3 respectively. For both test collections, document vector lengths do not have a significant effect on performance of the system. This may be due to insensitivity of the method to the parameters. Even a small number of terms is enough to compete with full length document vectors. For NEDT, Can et al. [9] found that using all terms in cosine similarity calculation gives better results. With all vector length configurations, cosine similarity-based ND method outperforms baseline significantly in terms of statistical tests ($p \ll 0.001$).

In this method again results for optimistic test collection are higher. This is because of the appropriateness of the method for a less strict novelty definition. Zhang et al. [48] also has similar observations that their methods model a less strict redundancy definition better. Since there is not a significant difference between any of the vector lengths, we will be using "ALL" configuration in the following experiments to make them more reliable. However, less number of terms can also be used for efficiency. Of course, as we stated these results may be result of the method's insensitivity to the parameters. Using very small number of terms for calculation would not be expected to be reliable.

5.2.1.3 Language Model-based ND Results

We experimented with two different smoothing algorithms in language model-based ND. Results of these are given in Table 5.4. Both of the algorithms give similar performances. Shrinkage smoothing has more smoothing power and ideally has the ability to approximate probabilities more accurately, so we would expect Shrinkage to outperform Dirichlet smoothing in both test collections but the results are consistent with both [5, 48]. In both of these studies, Shrinkage and Dirichlet smoothing methods have similar performance values. Language model-based ND method also outperforms baseline significantly in terms of statistical

Table 5.2: Average results of cosine similarity-based ND method with optimistic test collection with varying document vector lengths.

	Training			**Test**		
Doc. Vec. Length	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
10	0.780	0.938	0.848	0.782	0.938	0.848
20	0.770	0.954	0.848	0.770	0.949	0.845
30	0.770	0.961	0.851	0.773	0.961	0.853
40	0.775	0.959	0.854	0.771	0.954	0.848
50	0.777	0.962	0.856	0.776	0.960	0.854
60	0.778	0.961	0.856	0.774	0.957	0.852
70	0.805	0.929	0.858	0.803	0.931	0.858
80	0.800	0.935	0.858	0.800	0.933	0.856
90	0.803	0.936	0.860	0.802	0.937	0.860
100	0.801	0.935	0.858	0.802	0.936	0.859
110	0.798	0.937	0.857	0.797	0.937	0.857
120	0.782	0.961	0.858	0.777	0.949	0.850
130	0.786	0.956	0.858	0.783	0.946	0.852
140	0.776	0.971	0.858	0.775	0.971	0.858
150	0.777	0.967	0.858	0.774	0.963	0.854
160	0.777	0.969	0.858	0.778	0.965	0.857
170	0.782	0.960	0.858	0.773	0.954	0.850
180	0.777	0.966	0.857	0.775	0.959	0.853
190	0.775	0.968	0.857	0.776	0.965	0.856
200	0.780	0.961	0.857	0.775	0.953	0.851
ALL	0.778	0.963	0.857	0.776	0.954	0.852

Table 5.3: Average results of cosine similarity-based ND method with pessimistic test collection with varying document vector lengths.

	Training			**Test**		
Doc. Vec. Length	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
10	0.587	0.958	0.717	0.583	0.957	0.713
20	0.713	0.713	0.713	0.713	0.713	0.713
30	0.635	0.896	0.730	0.634	0.897	0.730
40	0.638	0.908	0.737	0.634	0.910	0.734
50	0.635	0.909	0.735	0.637	0.902	0.733
60	0.666	0.863	0.739	0.664	0.861	0.737
70	0.668	0.870	0.743	0.670	0.868	0.743
80	0.665	0.868	0.739	0.662	0.866	0.737
90	0.638	0.916	0.740	0.632	0.906	0.733
100	0.657	0.885	0.741	0.648	0.882	0.734
110	0.646	0.905	0.741	0.633	0.899	0.731
120	0.656	0.886	0.741	0.654	0.888	0.740
130	0.653	0.888	0.740	0.653	0.885	0.739
140	0.656	0.884	0.741	0.649	0.885	0.735
150	0.631	0.939	0.743	0.628	0.940	0.741
160	0.634	0.935	0.743	0.631	0.924	0.738
170	0.641	0.919	0.742	0.629	0.903	0.729
180	0.639	0.922	0.742	0.627	0.909	0.729
190	0.636	0.927	0.742	0.632	0.917	0.736
200	0.639	0.920	0.742	0.635	0.911	0.736
ALL	0.630	0.935	0.741	0.631	0.923	0.738

Table 5.4: Results of language model-based ND method.

Method	Ground Truth	Training			Test		
		Prec.	**Rec.**	**F-Mea.**	**Pre.**	**Rec.**	**F-Mea.**
Dirichlet	Pes.	0.747	0.904	0.806	0.741	0.900	0.801
	Opt.	0.859	0.929	0.890	0.859	0.930	0.889
Shrinkage	Pes.	0.750	0.892	0.802	0.744	0.887	0.796
	Opt.	0.841	0.942	0.885	0.838	0.933	0.880

tests ($p \ll 0.001$).

5.2.1.4 Cover Coefficient-based ND Results

In this section, we provide the results of the cover-coefficient based ND method and compare it with best configurations of the previously presented results. Table 5.5 shows the results. Best performing method amongst all of the methods is language model-based ND with Dirichlet smoothing which outperforms the other proposed methods significantly ($p \leq 0.002$). This observation is generally consistent with ND studies conducted in English. As we also stated before, KL divergence is an appropriate measure for novelty because of its asymmetry. Biggest issue in language models is the smoothing and it seems that Dirichlet smoothing may satisfy the needs. It is easy to calculate and does not require any reference collection for smoothing.

Second best performing system, Cosine similarity-based ND is also one of the best performers in ND studies in English. Using different vector lengths, we also showed that in some applications this method may not be sensitive to parameters. Additionally, it is not necessary to use complex term weighting functions because cosine similarity also works well with raw term frequencies. Even these frequencies are not required to be normalized because cosine similarity has its own normalization mechanism [4]. As we stated before, a shorter document vector length may be chosen if efficiency is a real issue in the task, however if a stopword list is utilized in the study, it is not generally necessary to shorten the

Table 5.5: Results of all methods' best configurations.

Method	**Ground Truth**	**Training**			**Test**		
		Pre.	Rec.	F-Mea.	Prec.	Rec.	F-Mea.
CC	Pes.	0.550	0.928	0.681	0.542	0.923	0.672
	Opt.	0.689	0.980	0.806	0.686	0.973	0.801
LM-Dirichlet	Pes.	0.747	0.904	0.806	0.741	0.900	0.801
	Opt.	0.859	0.929	0.890	0.859	0.930	0.889
Cosine-All	Pes.	0.630	0.935	0.741	0.631	0.923	0.738
	Opt.	0.778	0.963	0.857	0.776	0.954	0.852
Random	Pes.	*No training result*			0.498	0.500	0.491
	Opt.				0.678	0.500	0.573

document vector length.

Cover coefficient as the least effective proposed method outperforms random baseline significantly in terms of statistical tests ($p \ll 0.001$) in both of the ground truth types. When compared to language model method, superiority of cover coefficient based-ND method is that it only has one parameter.

5.2.1.5 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 it can be seen in Table 5.6, there is no significant difference between the performances obtained by category-based threshold learning and general learning. Although there is no significant difference, these results are promising that 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 13 categories, some categories may have very few topics such as 3. Even if we apply leave-one-out cross validation, the data size may still not be enough to learn a threshold value accurately. Category (or broader topics) were studied in topic detection content also in TREC event and opinion type topics were studied differently by some researchers but this type of category information was not utilized before. These results show that category

Table 5.6: Results of best performances of each system with general and category-based threshold learning.

Method	Ground Truth	General	Category
Cover	Pessimistic	0.671	0.676
	Optimistic	0.802	0.800
Cosine	Pessimistic	0.737	0.734
	Optimistic	0.853	0.851
Language	Pessimistic	0.800	0.796
	Optimistic	0.889	0.883

information usage should be examined further.

5.2.1.6 Comparison of the Proposed Methods on An Example

In this section, we compare our proposed methods on a toy collection of documents taken from topic 1 in Table A.1 which is about a traffic accident in Kars, Turkey. We used the first story and the following six documents for the comparisons. Documents are given in Appendix B. Documents are in chronological order from document 1 to document 7 where document 1 is the first story. Table 5.7 shows the novelty values obtained by the methods between each document. For cosine similarity-based and CC-based methods, novelty measure is a similarity value where for LM-based method it is a distance measure (KL divergence). Values in the table are obtained by using the optimal parameters obtained from training. First two columns show the document numbers which are being compared. Following three columns give the novelty values for cosine, CC and LM based ND methods respectively. The last column indicates the novelty decision for the document, 0 stands for not novel and 1 stands for novel (same decisions are given by both of the annotators).

As we described in Section 3, our methods calculate some novelty measure between an incoming document and all of the previous documents. Then, if all of these values do not fail comparison condition with the threshold value, document is labeled as novel, otherwise it is labeled as not novel. For cosine similarity-based

Table 5.7: Novelty measure values obtained for each proposed method between the documents in the toy collection.

Document 1	Document 2	Cosine	CC	LM	Novelty
1	0	0.43	0.56	7.32	0
2	0	0.46	0.55	5.42	0
2	1	0.60	0.12	4.77	0
3	0	0.46	0.46	6.50	1
3	1	0.51	0.08	6.39	1
3	2	0.45	0.08	6.25	1
4	0	0.41	0.44	6.59	1
4	1	0.33	0.05	7.30	1
4	2	0.46	0.06	6.69	1
4	3	0.35	0.08	6.95	1
5	0	0.45	0.36	4.95	0
5	1	0.64	0.08	4.45	0
5	2	0.62	0.06	4.32	0
5	3	0.59	0.10	4.49	0
5	4	0.56	0.10	3.33	0
6	0	0.44	0.36	6.69	0
6	1	0.59	0.06	6.56	0
6	2	0.50	0.04	6.64	0
6	3	0.45	0.06	7.12	0
6	4	0.41	0.06	6.25	0
6	5	0.63	0.11	5.78	0

method, if we use 0.52 threshold, all novel documents will be detected where there will also be a false positive, document 1. In CC-based method, when threshold is taken as 0.47 there will be two false positives. Also, LM-based method will have one false positive when threshold is taken as 5.77. These values show that with a static threshold value, even if we pick the threshold value by hand, there will be mistakes.

5.2.2 TREC Novelty Track 2004 Results

In order to evaluate effectiveness of our methods in different languages and test collections, we experimented with TREC 2004 test collection. As we mentioned

Table 5.8: Test results for of cover coefficient-based ND method and 5 participants of TREC 2004.

Participant *(Run Name)*	**Precision**	**Recall**	**F-Measure**
Dublin City U. *(CDVP4nterf1)*	0.4904	0.9038	0.6217
Meiji U. *(MeijiHIL2WRS)*	0.4790	0.9310	0.6188
U. of Mass. Amherst *(CIIRT2R2)*	0.4712	0.9544	0.6176
omitted results			
C. for Computer Science *(ccsmmr5t2)*	0.4326	0.9938	0.5880
Cover Coefficient	0.4334	1.0000	0.5867
Meiji U. *(MeijiHIL2CS)*	0.4246	0.9952	0.5797

in 4.2, TREC Novelty 2003 data is also used for training. We only run cover coefficient-based ND method on TREC 2004 data since both cosine similarity and language models were used in the track by other participants.

The results can be seen in table 5.8. Since there were many participants we only included results of five runs from Task 2. First three rows show the best performing three systems of Task 2. The important result here is *CIIRT2R2* because they use cosine similarity for ND. This finding is similar to our findings in BilNov that cosine similarity-based ND method outperforms cover coefficient-based method [19]. Additionally, in their previous study Allan et al. [5] shows that language model-based ND methods outperform cosine similarity-based method in TREC 2003 data. When all of these results are examined, we can assert that results are consistent with the results in Turkish.

Cover coefficient-based ND outperforms the baseline in Task 2 and ranks 35. within 55 participants. This may be counted as a promising result since 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 sentences. Additionally, a complex threshold mechanism can be employed.

Chapter 6

Conclusion & Future Work

In this study we presented our findings on ND in Turkish. ND problem was not previously studied in Turkish and so we built a Turkish ND test collection which contains 59 topics with an average of 51 tracking documents. We presented statistics about this test collection. We proposed usage of three ND methods, a cosine similarity-based method, a language model-based method and a cover coefficient-based method where first two are motivated from the previous studies on ND. We showed that usage of different document vector lengths for cosine similarity calculation does not have a significant effect on the system performances. Additionally, for language model-based ND method, we showed that a simpler smoothing method, Dirichlet smoothing, can have similar performance with a more complex smoothing method, Shrinkage smoothing. In addition to these two methods, we proposed cover coefficient-based ND method. We also proposed a random baseline for ND which was not used before. Also, first time in the ND context, we experimented on category-based threshold learning which uses topics from the same category when learning a threshold. This was motivated by the differences between characteristics of news from different categories. Although, the results of category-based and general threshold learning do not report any significant difference, it is promising to see even with a small set of topics from the same category, learning process can be conducted without decreasing the performance. Finally, we provided results of cover coefficient-based ND method in

TREC 2004 Novelty test collection. Cover coefficient-based ND method ranks 35 in 55 participants. The results are promising but there should be done some adaptations on the method.

Although ND was studied in information retrieval for three years in TREC Novelty Tracks, there are still a lot to do in both information retrieval and other domains. This study was one of the first studies to apply ND on tracking documents of a topic. 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.

Some future pointers for ND studies are :

1. We need to utilize category information in a more complex way and evaluate this with an appropriate test collection which contains plenty of topics per category.

2. When working on documents, instead of considering documents as a whole, sentences may be processed separately. Additionally, some of the sentences in a document can be irrelevant and may contain novel information. These type of sentences may be eliminated before application of ND. For an evaluation of sentence level relevance detection, TREC Novelty Track test collection may used or a new test collection may be created as well.

Bibliography

[1] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. J. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling: report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, 2002. *SIGIR Forum*, 37(1):31–47, 2003.

[2] J. Allan, J. Carbonell, G. Doddington, and J. Yamron. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[3] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18, New York, NY, USA, 2001. ACM.

[4] J. Allan, V. Lavrenko, and R. Swan. Explorations within topic tracking and detection. In *Topic detection and tracking: event-based information organization*, pages 197–224. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[5] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *SIGIR '03: Proceedings of the 26th Annual International*

ACM SIGIR Conference on Research and Development in Informaion Retrieval, pages 314–321, New York, NY, USA, 2003. ACM.

[6] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[7] S. Blott, F. Camous, P. Ferguson, G. Gaughan, C. Gurrin, G. J. F. Jones, N. Murphy, N. E. O'Connor, A. F. Smeaton, P. Wilkins, O. Boydell, and B. Smyth. Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[8] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar. Bilkent news portal: a personalizable system with new event detection and tracking capabilities. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in lnformation Retrieval*, pages 885–885, New York, NY, USA, 2008. ACM.

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

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

[11] F. Can and E. A. Ozkarahan. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Trans. Database Syst.*, 15(4):483–517, 1990.

[12] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[13] J. G. Conrad and C. P. Schriber. 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, 2006.

[14] J. M. Conroy. A hidden markov model for the trec novelty task. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[15] D. Eichmann, Y. Zhang, S. Bradshaw, X. Y. Qiu, L. Zhou, P. Srinivasan, A. K. Sehgal, and H. Wong. Novelty, question answering and genomics: The University of Iowa response. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[16] D. Harman. Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2002.

[17] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

[18] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[19] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at trec 2004: Novelty and hard. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[20] F. Jelinek and R. L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, 1980.

[21] S. Kardas. New event detection and tracking in Turkish. Master's thesis, Department of Computer Engineering, Bilkent University, 2009.

[22] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304, New York, NY, USA, 2004. ACM.

[23] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[24] L. S. Larkey, J. Allan, M. E. Connell, A. Bolivar, and C. Wade. Umass at trec 2002: Cross language and novelty tracks. In *Proceedings of the Eleventh Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2002.

[25] G. L. Lewis. *Turkish Grammar*. Clarendon Press, Oxford, 1967.

[26] X. Li and W. B. Croft. An information-pattern-based approach to novelty detection. *Inf. Process. Manage.*, 44(3):1159–1188, 2008.

[27] M. Markou and S. Singh. Novelty detection: a review–part 1: statistical approaches. *Signal Processing*, 83(12):2481 – 2497, 2003.

[28] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 280–285, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[29] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 198–207, New York, NY, USA, 2005. ACM.

[30] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, Nov 1996.

[31] H. C. Ocalan. Bilkent news portal: A system with new event detection and tracking capabilities. Master's thesis, Department of Computer Engineering, Bilkent University, 2009.

[32] R. Papka. *On-line New Event Detection, Clustering and Tracking*. PhD thesis, Department of Computer Science, University of Massachusetts, 1999.

[33] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York, NY, USA, 1998. ACM.

[34] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, and R. S. Raghavan. Newsinessence: a system for domain-independent, real-time news clustering and multi-document summarization. In *HLT '01: Proceedings of the First International Conference on Human Language Technology Research*, pages 1–4, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[35] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *Proceedings of the Third Text REtrieval Conference, TREC*, 1995.

[36] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[37] I. Soboroff. Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[38] I. Soboroff and D. Harman. Overview of the TREC 2003 novelty track. In *Proceedings of the Twentieth Text REtrieval Conference, TREC*, pages 38–53, Gaithersburg, MD, USA, 2003.

[39] I. Soboroff and D. Harman. Novelty detection: the TREC experience. In *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[40] M.-F. Tsai, M.-H. Hsu, and H.-H. Chen. Similarity computation in novelty detection and biomedical text categorization. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[41] E. Uyar. Near-duplicate news detection using named entities. Master's thesis, Department of Computer Engineering, Bilkent University, 2009.

[42] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[43] H. R. Varian. Universal access to information. *Commun. ACM*, 48(10):65–66, 2005.

[44] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR '98: Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in lnformation Retrieval*, pages 28–36, New York, NY, USA, 1998. ACM.

[45] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *KDD '02: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 688–693, New York, NY, USA, 2002. ACM.

[46] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[47] H. Zhang, H. Xu, S. Bai, B. Wang, and X. Cheng. Experiments in trec 2004 novelty track at cas-ict. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*, Gaithersburg, MD, USA, 2004.

[48] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in lnformation Retrieval*, pages 81–88, New York, NY, USA, 2002. ACM.

Appendix A

Turkish ND Test Collection Topics

Table A.1: BilNov statistics.

Num. : Topic (BilCol2005 Num)	# of Docs.	Avg. Nov. Ratio	Avg. Time
1 : Kars'ta trafik kazası (1)	20	57.5	30
2 : Onur Air'in Avrupa'da yasaklanması (2)	80	61.25	105

Continued on Next Page...

Table A.1 – Continued

Num. : Topic (BilCol2005 Num)	# of Docs.	Avg. Nov. Ratio	Avg. Time
3 : Nema karşılığı kredi (4)	31	72.59	35
4 : Londra metrosunda patlama (6)	80	43.13	50
5 : Çocuk tacizi skandalı (7)	80	67.5	72.5
6 : Formula G (8)	20	70	20
7 : Şemdinli olayları (11)	79	66.25	
8 : Türkiye'de kuş gribi (12)	80	53.75	162.5
9 : Fenerbahçe'nin şampiyon olması (13)	80	65.63	87.5
10 : Mortgage Türkiye'de (14)	80	63.13	107.5
11 : 2005 Avrupa Basketbol Şampiyonası (15)	78	53.85	67.5
12 : Van Yüzüncü Yıl Üniversitesi Rektörü Prof.Dr. Yücel Aşkın'ın tutuklanması (16)	80	60	155
13 : Kral Fahd'ın hastaneye kaldırılması (17)	51	66.67	60
14 : Memurlarının bir üst dereceye çıkması (18)	52	50	57.5
15 : Bill Gates'in Türkiye'ye gelmesi (19)	17	73.53	25
16 : Mısır'da üst üste patlamalar (20)	80	50.63	120
17 : Atilla İlhan'ın vefat etmesi (21)	40	66.25	45
18 : Ata Türk'ün öldürülmesi (22)	43	56.98	87.5

Continued on Next Page...

Table A.1 – Continued

Num. : Topic (BilCol2005 Num)	# of Docs.	Avg. Nov. Ratio	Avg. Time
19 : DT Genel Müdürünün görevden alınması (23)	63	75.4	52.5
20 : Universiade 2005 (24)	80	85	62.5
21 : Yahya Murat Demirel'in Bulgaristan'da yakalanması (25)	80	55.63	67.5
22 : Bağdat El Ayma Köprüsü üzerinde izdihamda çok sayıda insanın ölmesi (26)	29	50	30
23 : Prof. Dr. Sadettin Güner ve oğlunun Trabzon'da öldürülmesi (27)	41	62.2	50
24 : Nermin Erbakan'ın tedavi altına alınması (29)	45	57.78	67.5
25 : 15. Akdeniz Oyunları (31)	80	73.13	95
26 : Kemal Derviş'in UNDP Başkanı seçilmesi (32)	80	45	72.5
27 : Caferi'nin tarihi Tahran ziyareti (33)	22	72.73	27.5
28 : Gediz'de grizu patlaması (34)	39	58.97	42.5
29 : Sarıgül'ün kendini savunması (35)	80	55	80
30 : Paris'te göstericilerin polisle çatışması (36)	80	57.5	75
31 : 2005 Nobel Tıp Ödülü gastrit ve ülserin bakterilerden kaynaklanması (39)	19	50	20
32 : Kayseri Erciyes Üniversitesi bebek ölümleri (40)	39	58.98	35

Continued on Next Page...

Table A.1 – Continued

Num. : Topic (BilCol2005 Num)	# of Docs.	Avg. Nov. Ratio	Avg. Time
33 : Marburg virüsünden ölenler (41)	25	64	15
34 : Gamze Özçelik'in görüntülerinin internette yayınlanması (42)	43	66.28	37.5
35 : Türkiye'nin ilk yediz bebekleri (43)	56	65.18	90
36 : Yeni Türk Ceza Kanunu'nun yürürlüğe girmesi (44)	53	64.15	105
37 : Saddam Hüseyin'in yargılanmaya başlanması (45)	80	55	77.5
38 : Beylikdüzü'nde çöpte patlama (46)	17	52.94	20
39 : Endonezya'nın Bali Adası'nda eşzamanlı patlamalar (47)	15	46.67	22.5
40 : Sahte rakı (48)	80	50.63	82.5
41 : Hindistan'da meydana gelen patlamalar (49)	21	78.57	27.5
42 : Bülent Ersoy ve Deniz Baykal polemiği (50)	52	51.93	42.5
43 : Sochi seferini yapan Ufuk-1 gemisinin yanması (52)	20	57.5	20
44 : İstanbul'da Dünya Kadınlar Günü için gösteri yapanları coplayan 3 polisin açığa alınması (54)	80	54.38	87.5
45 : Kuşadası'nda minibüsde patlama (55)	50	41	115
46 : Esenboğa Havalimanı İç Hatlar Terminali'nin yanması (56)	18	55.56	20
47 : Zeytinburnu'nda bir evde patlama (57)	28	44.65	27.5

Continued on Next Page...

Table A.1 – Continued

Num. : Topic (BilCol2005 Num)	# of Docs.	Avg. Nov. Ratio	Avg. Time
48 : Malatya Çocuk Yuvası'nda işkence (58)	80	68.75	97.5
49 : Prof Dr. Kalaycı'ya Silahlı Saldırı (60)	44	48.87	47.5
50 : 15 Yeni Üniversite Kuruluyor (62)	59	46.61	45
51 : Gaziantep'te Tanker Patlaması (63)	33	56.07	27.5
52 : Kzım Koyuncunun Ölümü (66)	30	70	40
53 : Melih Kibar ın Ölümü (67)	16	68.75	22.5
54 : Japonya Osaka'da Tren Kazası (71)	29	60.35	32.5
55 : Yunanistan'da Türk Bayrağına Çirkin Saldırı (74)	55	40	77.5
56 : Maslak'ta Patlama (75)	30	56.67	37.5
57 : Rum Yolcu Uçağının Düşmesi (77)	80	56.25	90
58 : Zeytinburnu'nda Geminin Batması (79)	38	44.74	52.5
59 : Bin Yıllık Yolculuk Sergisi (80)	22	50	37.5
Average	50.9	59.36	60.17

Appendix B

Toy Test Collection

Document 1 28.05.2005 08:50:00. Otobüs Aras Nehri'ne uçtu: 7 ölü, 7 kayıp, 38 yaralı.. Cem Bakırcı, Onur Sağsöz, Menderes Uray, Mukadder Yardımcıel Sarıkamış (Kars),(DHA) İstanbul'dan Iğdır'a giden ve içinde 49 kişinin bulunduğu otobüs, Kars'ın Sarıkamış ilçesi yakınlarında sürücünün yola düşen kayaya çarpmamak için direksiyon kırması sonucu Aras Nehri'ne uçtu. Yolcularının çoğunu Nahcivan'a giden Azerilerin oluşturduğu otobüstekilerden 7'si öldü, 38'i de yaralı kurtarıldı. Otobüsteki 7 kişiden henüz haber alınamazken bazılarının nehir sularında kaybolduğu sanılıyor. Iğdırlı tur şirketine ait 34 YJ 9924 plakalı otobüsün sürücüsü 50 yaşındaki Musa Telek, bu sabah saat 06:45 sıralarında Sarıkamış'ın Karakurt bucağı yakınlarına geldiğinde, heyelan nedeniyle yola düşen kayaya çarpmamak için direksiyon kırdı. Stabilize yolda aşırı hız nedeniyle sürücünün kontrolünden çıkan otobüs, yol kenarında bulunan Aras Nehri'ne uçtu. Horasan'a 44, Karakut'a 4 kilometre uzaklıkta meydana gelen kazada, otobüs nehirin ortasına yan yatarken, 46 yolcu ve 3 personel olmak üzere otobüsteki 49 kişi kendilerini nehir sularında buldu. Çoğu uykuda olan yolcular, can pazarında boğulmaktan kurtulmak için çırpınmaya başladı. Kazanın şokundan kurtulan yolcular kendilerini kıyıya atarken, imdat çığlıkları atan diğer yolcuları kurtarmak için yeniden nehre girdiler. Bu arada bazı yolcular kıyıya çıkamayarak nehrin azgın sularıyla sürüklenip kayboldu. Yoldan geçen araçlardan inenler, sudaki yolcuların çıkarılmasına yardım etti. Kurtarma

ekipleri olay yerinde kazanın duyulması üzerine bölgeye jandarma birlikleri, acil servis görevlileri ve ambulanslarla, erzurum sivil savunma birlik müdürlüğü'nden 5'i dalgıç 14 personel sevkedildi. Sudan çıkarılan ve yolda bekleşen yaralılar, ambulanslarla sarıkamış, kars, horasan ve erzurum'daki hastanelere nakledildi. Aras Nehri'nde arama çalışması başlatan ekipler, otobüste sıkışan Fatma Tuna'yı kurtardı ve kimlikleri saptanamayan 3 kişinin cesedini çıkardı. Erzurum Numune Hastanesi'nde 13'ü Azeri 17, Erzurum Aziziye Araştırma Hastanesi'nde 4'ü Azeri 7, Sarıkamış Devlet Hastanesi'nde 8'i Azeri 10, Kars Devlet Hastanesi'nde ise 4 Azeri olmak üzere toplam 29'u Azeri 38 yolcu tedavi altına alındı. Numune Hastanesi'nde tedavi altına alınan 54 yaşındaki Azerbaycanlı Gülşen Abdulleyev'in bir ara kalbi durdu. Doktorlar kalp masajı yaparak Gülşen Abdulleyev'i hayata döndürmeyi başardı. Erzurum kriz merkezi otobüs sürücüsü Musa Telek ile birlikte 7 kişiden henüz haber alınamadığını bildirdi. Kaza yerine gelen ve kurtarma çalışmalarını izleyen Kars valisi Nevzat Turhan, henüz kesin olmamakla birlikte otobüste 49 kişinin bulunduğunu açıkladı. Otobüs firmalarının sürücü seçiminde gerekli hassasiyeti göstermediklerini söyleyen vali Turhan, karayollarında kontrol yapılmayan bölgelerde hız sınırının aşıldığını curguladı. Şoför ve personel kaçtı mı? Vali Turhan, haber alınamayan 7 kişiden bazılarının Aras Nehri'nin bulanık sularında kaybolduğunun sanıldığını bildirdi. Otobüs sürücüsü Musa Telek, ikinci şoför ve muavinin kazadan yara almadan kurtuldukları ve olay yerinden kaçtıkları öne sürüldü. Şoför Musa Telek'in 1983 yılında ehliyet aldığı ve trafik kayıtlarına göre 11 kez hız sınırını aştığı için ceza kesildiği belirlendi. Yaralı anlatıyor kazadan hafif yaralı kurtulan Azeri Hüseyin Mehmetov, Horasan'da mola verdik ve yola çıktık. Sabahın erken saatleriydi. Otobüs vadide hızla gidiyordu. Aniden yol ortasındaki kayayı gören şoför, direksiyon kırdı ama kontrolü kaybetti. Yoldan nehire düştük. Bu sırada uyuyan yolcular uyandı. Ortalık ana baba gününe döndü. Yolculardan bazıları takla attığı sırada çevreye savruldular. İki çocuk vardı otobüste, birini kurtardılar ama diğerini göremedim dedi.

Document 2 28.05.2005 08:52:00. Otobüs, Aras Nehri'ne uçtu :21 yaralı. Otobüs, Aras Nehri'ne uçtu :21 yaralı. Bir yolcu otobüsünün Aras Nehri'ne uçması sonucu ilk belirlemelere göre 21 kişi yaralandı. İstanbul'dan Iğdır'a gitmekte olan Musa Pelek'in kullandığı 34 YJ 9924 plakalı yolcu otobüsü, Karakurt

mevkiinde Aras Nehri'ne uçtu. Olayda, ilk belirlemelere göre 21 kişi yaralandı. Otobüste bulunan yolcuların kurtarılmasına çalışılıyor.

Document 3 28.05.2005 09:06:00. Yolcu otobüsü Aras Nehri'ne uçtu. Sabah saat 06:00 sıralarında meydana gelen kazada, İstanbul'dan Iğdır'a gitmekte olan Musa Pelek'in kullandığı 34 YJ 9924 plakalı yolcu otobüsü, Karakurt mevkiinde Aras Nehri'ne uçtu. Yaralı yolcular Horasan ve Sarıkamış Devlet Hastaneleri'ne kaldırıldı. Olay yerine gelen vincin otobüsü çıkartması bekleniyor. Kazada ölen olup olmadığı konusunda henüz bir bilgi geçilmedi.

Document 4 28.05.2005 09:30:00. Aras Nehri'ne uçan otobüsten yaralı kurtarılan kadın öldü. Aras Nehri'ne uçan otobüsten yaralı kurtarılan kadın öldü Aras Nehri'ne uçan yolcu otobüsünden yaralı kurtarılan bir kadın, hayatını kaybetti. İstanbul'dan Iğdır'a gitmekte olan ve Karakurt mevkiinde Aras Nehri'ne uçan yolcu otobüsünden yaralı kurtarılan kadın, Sarıkamış Devlet Hastanesi'ne kaldırılırken öldü. Arama ve kurtarma çalışmalarına katılmak üzere, Erzurum Sivil Savunma Müdürlüğü'nden 5'i dalgıç 14 personelden oluşan bir ekip, olay yerine gitti. Otobüste bulunanların sayısı henüz tespit edilemedi. Arama ve kurtarma çalışmaları için sivil savunma ekibi bekleniyor.

Document 5 28.05.2005 10:38:00. Otobüs kazasında ölü sayısı 6 oldu. Erzurum'dan giden ve 5 dalgıcın yer aldığı sivil savunma ekibi arama çalışmalarını sürdürürken, kazada yaralananların Kars, Sarıkamış ve Horasan devlet hastanelerinde tedavileri sürüyor. Sabah saatlerinde meydana gelen kazada, Musa Pelek'in kullandığı bildirilen 34 YJ 9924 plakalı yolcu otobüsü, Karakurt mevkiinde Aras Nehri'ne uçmuştu. Edinilen bilgiye göre, olay yerinde 3 kişinin hayatını kaybettiği kazada, Erzurum'a sevkedilen yaralılardan Nahile Askerova da hayatını kaybetti. Kaza sonrası nehirde akıntıya kapıldığı belirtilen 3 kişiden 2'sinin cesedi yapılan çalışma sonrası bulunurken, diğer kayıp yolcunun aranmasına devam ediliyor.

Document 6 28.05.2005 10:38:00. Yolcu otobüsü Aras Nehri'ne uçtu :2 ölü. Yolcu otobüsü Aras Nehri'ne uçtu :2 ölü Aras Nehri'ne uçan yolcu otobüsünden, bir kadın yolcunun cesedi çıkarıldı. İstanbul'dan Iğdır'a gitmekte olan ve Karakurt mevkiinde Aras Nehri'ne uçan yolcu otobüsünden, dalgıçlar, bir kadın

yolcunun cesedini çıkardılar. Erzurum'dan giden ve 5 dalgıcın yer aldığı sivil savunma ekibi arama çalışmalarını sürdürürken, kazada yaralananların Kars, Sarıkamış ve Horasan devlet hastanelerinde tedavileri sürüyor. Sabah saatlerinde meydana gelen kazada, Musa Pelek'in kullandığı bildirilen 34 YJ 9924 plakalı yolcu otobüsü, Karakurt mevkiinde Aras Nehri'ne uçmuş, ilk belirlemelere göre 1 kadın, hastaneye götürülürken yaşamını yitirmiş, 20 kişi yaralanmıştı. Çıkarılan kadın yolcuyla birlikte, ölenlerin sayısı 2'ye yükseldi.

Document 7 28.05.2005 11:04:00. Yolcu otobüsü Aras Nehri'ne uçtu :2 ölü, 37 yaralı. Yolcu otobüsü Aras Nehri'ne uçtu :2 ölü, 37 yaralı Kars'ın Sarıkamış İlçesi Kaymakamı Bayram Gale, Aras Nehri'ne uçan otobüste belirlemelere göre iki kişinin öldüğü, 37 kişinin yaralandığını bildirdi. Gale, bir yolcu otobüsünün sabah saatlerinde Aras Nehri'ne uçtuğunu belirterek, "Şu ana kadar yaptığımız belirlemelere göre kazada 2 kişi yaşamını yitirdi" diye konuştu. Kazada yaralanan 37 kişiden 4'nün Kars, 10'nun Sarıkamış Devlet Hastaneleri'nde tedavi altına alındığı, diğer yaralıların ise Erzurum ve Horasan'a sevk edildiği kaydedildi.