Use of place information for improved event tracking

Yun Jin a, Sung Hyon Myaeng b,*, Yuchul Jung b

a Chungnam National University, Daejeon 305-764, South Korea
b Information and Communications University, Daejeon 305-732, South Korea

Received 27 May 2006; accepted 25 July 2006

Abstract

The main purpose of topic detection and tracking (TDT) is to detect, group, and organize newspaper articles reporting on the same event. Since an event is a reported occurrence at a specific time and place and the unavoidable consequences, TDT can benefit from an explicit use of time and place information. In this work, we focused on place information, using time information as in the previous research. News articles were analyzed for their characteristics of place information, and a new topic tracking method was proposed to incorporate the analysis results on place information. Experiments show that appropriate use of place information extracted automatically from news articles indeed helps event tracking that identify news articles reporting on the same events.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Place information; Event tracking; TDT; News article

1. Introduction

It is time consuming and laborious for individuals to detect a new event and track news articles reporting on the particular event from a variety of resources (Allan, 2002; Nallapati, Feng, Peng, & Allan, 2004). Topic detection and tracking (TDT) (Allan, Carbonell, Doddington, Yamron, & Yang, 1998) attempts to automate the process and provide a topic that consists of a seminal event or activity, along with all directly related events or activities. Here an event is a reported occurrence at a specific time and place and the unavoidable consequences, whereas an activity is a connected set of actions that have a common focus or purpose (Yang, Pierce, & Carbonell, 1998). Unlike traditional information retrieval or filtering, TDT focuses not just on topicality but on events or activities which often occur at a specific time and specific place (Allan, 2000; Nallapati et al., 2004).

Time information has been used in TDT. Based on the observation that articles reporting on a particular event tend to appear within a time window of two weeks, publication dates of news articles were used in topicality-based document clustering (Mani, Schiffman, & Zhang, 2003). More recent approaches attempted to

* Corresponding author.
E-mail addresses: myaeng@icu.ac.kr, wkim@cnu.ac.kr (S.H. Myaeng).

0306-4573/S - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.007
automatically extract time information embedded in news article text and use it in determining whether the event described in an article refers to the event being tracked (Kim et al., 2004; Mani et al., 2000).

Place information seems to be as important as time in identifying an event, but it has neither been used nor tested extensively for usefulness in TDT. Possible reasons are:

- Time information such as publication dates is easily available from news articles, but place information needs to be extracted from news article text.
- Time information is one-dimensional and thus relatively easy to represent and compare, but place information is at least two-dimensional and more complex to determine its relationship with others. If place information is represented in a hierarchical way, for example, we should consider not only a parent–child relationship (e.g., Korea–Seoul) but also a sibling relationship (Seoul–Incheon).
- Compared to time information processing, place names seem to be more difficult to process because of the variety of expressions and ambiguities (e.g., Washington can be a person name or a state name, and there are several places with the name Rome in world geography).

Since our current work focuses on the use of place information for TDT, not on the techniques for extracting place information, we use existing techniques and resources for place name recognition. For instance, we made use of some portal sites, such as Naver, Paran and ESRI. The sites provide services related to place name information, such as hierarchical information and/or location on a map for a given place name. This service is useful for determining the spatial relationship between two place names.

For effective use of place information, it is important to understand the characteristics in an event description. For instance, the place name as well as the time information tends to appear in the first sentence of a news article describing an event. By giving different weights for place names occurring in different parts of a new article, we can increase the chance of associating a place name with an event.

In sum, the contributions made in the research are as follows:

- We analyzed news articles to find important characteristics of how place information is used.
- We developed a novel event tracking method that incorporates the characteristics of place information.
- We demonstrated the value of place information in event tracking.

2. Related work

Research on distinguishing place and organization names as part of named entity recognition was reported in Message Understanding Conference (MUC)-6 (Sundheim et al., 1995). With the goal of better utilizing place information, a work on classifying such place information into categories like “city”, “country”, “region”, and “water” was reported (Chinchor, 1998; Chung, Lim, Hwang, & Jang, 2004; Lee & Lee, 2004). Categorization of place names helps determining whether two names refer to the same location. For instance, tagging a place name “Geneva, New York” with [location] and [location], respectively, is less useful than tagging them with [city] and [state], since the latter can tell the relationship between “Geneva, New York” and “New York” with the aid of a place entity hierarchy. In other words, for example, “the accident in the state of New York” and “the accident in Geneva” may be considered the same event with a detailed place name tagging.

A few papers addressed the issue of using place information to improve effectiveness of TDT. Smith (2002) used place information for tracking similar historical events in unstructured history documents, not news paper articles. Documents containing the same dates and places, which were identified by named entity recognition, were considered as describing the same historical event. Juha, Helena, and Marko (2004) extracted person names, places, temporal expressions, and general terms to form four vectors, each representing one of the four types. The similarity between two documents was computed based on four vectors that were treated

1 http://naver.com/.
2 http://cube.paran.com/.
with the equal weights in computing similarities and clustering documents. For place information, all the place names were used as element of the vector.

### 3. Characteristics of place information in news paper

We first analyzed news articles of Chosunilbo in 2002, the news paper with the largest subscribers in Korea, to understand the characteristics of place information in news paper articles reporting on events and activities. The purpose of this analysis was twofold: to figure out whether knowledge on places would help event tracking and to improve the way of treating place names as general terms in event tracking.

An event can be classified as either “short-term” or “long-term,” depending on the time span of the event although the division is not always clear-cut. Examples of short-term events are: “collapse of a dam in Spain,” “12th world AIDS Conference in Geneva, Switzerland.” Examples of long-term, on-going events are: “India begins nuclear testing,” and “national elections in the Philippines.” Characteristically, short-term events tend to have clear temporal boundaries with clear beginning and ending times, and hence their reports appear in a relatively short time window. Based on our analysis of manually classified news articles, short-term events tend to have a time span of less than 30 days whereas long-term events’ time spans are usually more than 30 days. Since long-term events tend to have unclear time boundaries, time information is less useful in identifying news articles reporting on the same event, making information other than time window, such as place information, can play an important role. Based on our preliminary study with the 96 events and the news articles reporting on them in the TDT2 collection, we found that 67% of the events were long-term. In other words, we felt that a significant number of news articles would benefit from place information because they have unclear time boundaries.

#### 3.1. Concentration on a short time window

News articles on a particular event begin to appear from the day it occurs and continue for a while, mostly within two weeks, except for on-going events. While this tendency has been utilized in TDT when time information is used (Allan, 2000, 2002; Juha et al., 2004), it can be used when place names are used for TDT because a place name for an event also tends to be concentrated within the time window.

Fig. 1 shows the distribution of a sample of Korean news articles related to twenty distinct events. It appears that a bunch of articles are published during the first five days. By the end of two weeks, the number of related articles becomes almost zero. Of course, the number of articles and the time span vary depending on the importance of the event.

The English corpus we used showed a similar phenomenon when we analyzed the news articles belonging to the “short-term” events among those collected for the 96 events. As shown in Fig. 2(a), only 12% of the events were reported beyond the first one week period after the event.

![Fig. 1. Distribution of articles related to an event along the time line.](image-url)
However, the news articles belonging to the “long-term” events showed very different characteristics as in Fig. 2(b). While 44% of the articles appear within 1.5 month window, a large proportion of them continue to appear within three month and even six month windows. Given that 67% of the events are “long-term”, a tracking system needs features other than time to decide whether two articles report on the same event.

3.2. Decreasing levels of detail

We observed that descriptions of place names tend to be maximally precise at the time of the event and become simpler often at a higher level of abstraction. For example, “fire on A Street, Albany, New York” is referred to as “fire on D street” (a simpler form) or “fire in Albany” (a simpler form at a higher level of abstraction) at a later time.

Based on our analysis of news articles, we found that only 69% of the articles contain complete descriptions of place names; 23% use partial or simpler descriptions and 8% have no place names.

In order to take advantage of this tendency, we should be able to form an equivalence class of place names for a particular event. For example, “fire in Albany” can be an exact match with an event description like “fire on D Street, Albany, New York” if the place name hierarchy verifies that Albany is a city in the state of New York. However, a description like “fire on Elm Street, Albany” should be interpreted as referring to a different event because the street names conflict each other in the same city.

3.3. Locations of place names

It was observed that a majority of place names for the main event occur at the beginning of news articles. This tendency is similar to the fact that the first sentence of an article contains the key information and serves as a summary sentence in automatic summarization. It is natural that place information is included in the key sentences because it is a key element in describing the main thrust of the event being reported. Place names occur in later sentences, too, but they usually refer to other related events rather than the main event being described in the article.

Fig. 3 depicts the tendency described above. As shown in the graph, the majority of place names occur at the beginning of articles. This tendency becomes useful information when we develop an algorithm for handling multiple occurrences of place names in a single article. That is, the place names appearing at the beginning part of an article would be more likely to be related to the main event of the article.

3.4. The number of place names in an article

The number of place name occurrences in an article is an indication of whether the article describes an event; the more occurrences of place names in an article, the less chance it describes an event. That is, an event article tends to have only a small number of place names because it would mention a place name for the event without making references to other places.
4. Topic tracking using place names

This section describes the core part of the proposed approach to event tracking. We first describe how place names were identified, extracted, and canonicalized, and then explain how place name-based similarity was computed between a target class and a news article. Finally, we show how the total score was computed between a class and a news article.

4.1. Place name extraction

Place names are not always extractable unambiguously because a word or phrase can be a place name, organization name, or a person name depending on the contexts. For example, “Washington” in English can refer to a person, a state, or a city. In Korean, the fact that there is no distinction between upper and lower cases makes it difficult to distinguish proper nouns from general nouns. Moreover, some place names are identical to verb stems, making it even more difficult to recognize them. For example, 부여 (Booyuh), a city name, can be the stem of a verb, 부여하다 (Booyuh–Hada) although their meanings are not related at all.

For Korean text, we used the named entity recognizer (NER) (Lee & Lee, 2004) borrowed from ETRI to automatically extract place names from Korean news papers. This NER is a hybrid system combining both pattern rules and statistical techniques, with the following 13 place name tags: COUNTRY, PROVINCE, CITY, CAPITALCITY, COUNTY, CONTINENT, MOUNTAIN, OCEAN, RIVER, ISLAND, TOPOGRAPHY, CANYON, and BAY. The NER tagger accepts a news article and tags each place name with one of the 13 place names. The performance of the NER tagger was known to be 81.8% in $F$-measure.

For English text, place name recognition was done with MITRE’s Alembic-workbench. This tool follows the tagging standard of MUC-5 and MUC-6 to produce named entity tags in XML. In this tool, person, place, and organization names are tagged as well as time information, with 85.2% accuracy in $F$-measure (David et al., 1997). Compared to the Korean NER tagger, place name tags are not so specific.

4.2. Canonicalization and place hierarchy

Once place names are recognized and tagged with the labels, at least two problems need to be handled before they are used for event tracking. Since the actual descriptions of a place name in news articles may change as time goes by, the extracted place names must be examined for their hierarchical relationships. In a Korean news article, for example, 가복동 (Gabokdong) is used as the name of the specific place for an accident, but 군산시 (Goonsan city), the name of the city that covers the specific place, may be used in a later article referring to the same accident. In this case, the hierarchical relationship between the two place names must be established so that they are treated as referring to the same region as far as the accident is concerned.
The second issue is related with abbreviated form of place names, which are used when a place name is mentioned multiple times or referred to in a later article for a known event. For example, a province name "Chonbuk" (Chonbuk) is a short form for "천송도" (Cholabukdo), and their identity must be established by consulting a dictionary. The same phenomenon exists in English as in state names (e.g., CA is for California) and person names (e.g., Mike for Michael). We call this process as canonicalization.

As a way to handle these issues, we utilized resources available on the Web. Some Web sites Naver, Paran in Korea provide an information service for various regions and locations in Korea. They accept a place name as a query and return its canonicalized form (full name) and its super-region and sub-regions, together with other information about the place such as restaurants, motels, banks, etc.

Basically, the same methodology was employed for canonicalization of English place names. We used the place name retrieval system available at ESRI to extract detailed information including the hierarchical structure. Place names such as city or country names extracted from a named entity tagger were used as queries to the site, and the HTML pages returned results were parsed to obtain the necessary information.

4.3. Similarity calculation using place names

Given canonicalized place names and hierarchically related names, the next step is to incorporate the place information into the event tracking process per se. A news article is compared with representations of various events and classified into the best matching category (Kim et al., 2004). Consequently, similarity calculation between two articles is the key process for event tracking. In this sub-section, we describe four similarity measures that take into account the four characteristics of place names in news articles mentioned in Section 3.

4.3.1. Similarity based on a place name hierarchy

News articles may use names at different levels of abstraction in a place name hierarchy. Assuming that the target event description has the fullest and most specific expression for the place where the event occurred, we attempt to compute the degree to which a place name appearing in a news article matches the target place name. The computation specified below can be seen as measuring semantic distance between two nodes in a concept hierarchy although our hierarchy consists of place names rather than concepts. Similarity between \( n_t \), the set of the target event place name units, and \( n_d \), the set of name units of a place name in a news article document, is defined as

\[
\text{Sim}_h(n_t, n_d) = \begin{cases} 
\frac{|n_t \cap n_d|}{|n_t|} & \text{if there is no conflict} \\
0 & \text{otherwise}
\end{cases}
\]

where \(|\cdot|\) is the size of a place name, i.e., the number of name units. For example, if the target event place name is “D Street in Albany, New York,” \( n_t = \text{“D Street”}, \text{“Albany”}, \text{“New York”} \) whereas \( n_d = \text{“Albany”} \) when the place name in a new article is “Albany”. The similarity between \( n_t \) and \( n_d \) becomes 1/3. While “Albany, New York” and “Albany” would be the same node in the place name hierarchy if the latter actually refers to the city in New York, the former has a higher similarity value of 2/3 according to the formula. We believe this reflects that “Albany” without the specific state name carries some uncertainty, resulting in a smaller similarity value. It should be noted that “conflict” means similarity becomes zero if \( n_d \) is not a subset of \( n_t \) or if the most specific description units of the target event and the place name in the article do not match. For example, “D Street in Albany, New York” does not match with “Elm Street, Albany, New York,” resulting in a conflict.

4.3.2. Using the position of place names

According to the observation in Section 3, the location of place names in news articles seems to determine whether the place name is actually associated with the main event being described in the article. We compute the degree of certainty that a place name refers to the main event in an article as follows:

\[
\text{Loc}(n_d) = 1 - \frac{\text{Pos}(n_d)}{|d|}
\]
where $|d|$ is the number of sentences in the news article where the place name appears and $\text{Pos}(n_d)$ is the position of the sentence containing the place name, which starts from 0. If a place name appears in the first sentence, for example, $\text{Pos}(n_d)$ is 0 and $\text{Loc}(n_d)$ becomes the maximum value being 1.

4.3.3. Calculating coherence of place names

We discriminate against a news article containing many place names that seem to suggest that it does not describe a specific event. We compute the degree of place name coherence with respect to an event as follows:

$$\text{Coh}(n_d) = \omega + \left(1 - \omega\right) \frac{\sum n_d}{\sum N_d}$$

where $\sum n_d$ is the occurrence count of the place name being considered and $\sum n_d$ is the occurrence count of all the place names in the article. $\omega$ is the constant determined by experiments. For the work reported in this article, we used 0.5 as the constant value based on preliminary analysis.

4.4. Combining three features

Given an event description and an article, evidence gathered from the hierarchical information, position feature, and coherence feature needs to be combined to determine the final similarity value. Since the place name similarity with hierarchical information is the most crucial for matching, we use it as the basis and try to adjust it with the two feature values to compute the final similarity value $\text{Sim}_{\text{place}}$. Following is the heuristics we used:

- If $\text{Sim}_h(n_t, n_d) = 1.0$ and $\text{Loc}(n_d) < 0.5$ and $\text{Coh}(n_d) < 0.55$
  - $\text{Sim}_{\text{place}}$ is reduced to 0.8
- If $0.5 < \text{Sim}_h(n_t, n_d) < 0.8$ and $\text{Loc}(n_d) > 0.7$ and $\text{Coh}(n_d) > 0.7$
  - $\text{Sim}_{\text{place}}$ is increased to 0.8
- Otherwise, $\text{Sim}_{\text{place}} = \text{Sim}_h(n_t, n_d)$

The first rule is based on our observation that even though the matching between two place names is perfect, there is a chance that the one in a news article may not be the place for the event if the location and coherence feature values are very low. In such a case, the overall similarity value for place names should be lowered. The second case is to boost the similarity value that is sufficiently high when the location and coherence feature values are high enough to provide a reasonable level of confidence that the place name is indeed specific to the event being described.

4.5. Final similarity

The final similarity value between a given event category and a news article can be obtained by combining the content similarity, $\text{Sim}_{\text{content}}$, and place based similarity, $\text{Sim}_{\text{place}}$, as follows:

$$\text{Sim}_{\text{final}}(t, d) = \alpha \times \text{Sim}_{\text{content}}(t, d) + (1 - \alpha) \times \text{Sim}_{\text{place}}(t, d), \quad \text{where} \quad \alpha = 0.6$$

$\text{Sim}_{\text{content}}$ was computed using the Cosine similarity measure in the vector space model. The news articles were processed with the standard indexing method including stemming, removal of stop words, and TFIDF to compute the weights of terms. Details are described below.

4.6. Baseline system

Our topic tracking system uses an online single pass clustering method because it fits the dynamic nature of the online tracking situation and faster than offline clustering methods. For similarity calculation between an event-representing news article $t$ and an incoming news article $d$, we chose to use the Cosine similarity measure (Formula (4)) based on TF * IDF with Okapi’s weighting scheme (Formula (5)) (Robertson, Walker, Jones, Hancock-Beaulieu, & Gaford, 1995):
\[
\text{Sim}(d_j, t) = \frac{d_j \cdot \tilde{q}}{|d_j| \times |\tilde{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,t}^2}}
\]

\[
w_{i,j} = \left( \frac{tf}{tf + 0.5 + 1.5 \times \frac{\text{docLength}}{\text{avgDocLength}}} \right) \times \left( \frac{\log(N/df)}{\log N} \right)
\]

where

tf: term frequency in the document;
df: the number of documents containing the term;
N: the number of documents in the collection;
docLength: term length of the document;
\text{avgDocLength}: the average document length in the collection.

For the processing of English news article text, we used Porter stemmer provided by Lemur tool (Paul & Callan, 2001) and eliminated stop words. In order to avoid the problem of automatically converting acronyms involving special characters like a period (e.g., “U.S.” to “us”), a manually constructed acronym list was used.

5. Evaluation with Korean news articles

In order to test our initial hypothesis that explicit use of place information would help improving the event tracking task and also to verify our schemes for using place information, we conducted a set of experiments using 110,000 Korean news articles published in 2002 by Chosunilbo, which has the largest number of subscribers in Korea.

Among the 70 topics constructed for this collection, we selected 20 that include place names and use them as events for which event tracking would be done, and other 50 topics were excluded. For each event, relevant articles were pre-selected by people and used for evaluation of system-generated output. The total number of “answer” articles is 990 for the 20 events. Relevance judgments were made by two people after reading the articles. Table 1 shows a sample of the event topics and the number of “answer” documents.

5.1. Evaluation of place name feature

Before we finalize our event tracking system with explicit place information, we tested the efficiency of individual features of the place names as reflected in the formulas (1)–(3) in Section 4. This set of experiments would help us finalize our method of combining pieces of evidence coming from different aspects of place information and test our initial hypotheses based on the characteristics of place names in news articles.

We came up with four different cases beyond the baseline that is just based on content similarity. The first three are for the three different formulas and the fourth one is for a combination of the three. In addition, we implemented the method used in Juha et al. (2004) to validate our approaches. Following are all the cases compared:

<table>
<thead>
<tr>
<th>Event topics</th>
<th>ID</th>
<th>Event title</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fire in the amusement quarter of Gunsan-si</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Protest against Kim Dong-sung’s disqualification at the Winter Olympics Game</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td>...</td>
<td>Death of Sohn Kee-chung a marathon hero</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>990</td>
</tr>
</tbody>
</table>
As shown in Table 2, the results are promising in that each feature appears to make some contribution and the combination of the three features gave a 7.27% of improvement over the baseline that does not use explicit place information. Given that it is more difficult to extract place information accurately than time information that is immediately available with publication dates, the results seem to show some potential. Even the relatively low performance increases in Cases C and D attest to their values when combined with others. It should be noted that the previous approach in Juha et al. (2004) gives performance increase almost identical to that of individual features as in the last row of the table. A conclusion we can draw from the experiments is that it is worthwhile using various features of place names based on their characteristics and take them into consideration when place information is used in combination with the content of articles.

5.2. Evaluation with place and time information

We ran experiments to investigate on the complementary aspects of place and time information. With two different types of time information, publication dates and those automatically extracted from text as in Kim et al. (2004), we investigated three different cases:

Case 1: use of both place and time information (extracted from text),
Case 2: use of publication dates (time information only),
Case 3: use of both place information and publication dates,
Case 4: use of both extracted time and publication dates (time information only),
Case 5: use all the three kinds of information.

As in Table 3, the result shows that when combined with place information, publication dates were slightly more useful than the time information extracted from the news paper text. Comparing Case 2 and Case 3,
adding place information improved effectiveness by about 2%, whereas the improvement from Case 4 to Case 5 shows a slightly less improvement. We feel that the amounts of improvement over the use of time information is not as great as expected, perhaps due to the inaccuracies of place name recognition. Nonetheless, by using all the three types of information as in Case 5, we obtained the best performance. This implies that although the improvements we get from individual sources may be small, their combination can make a significant improvement.

5.3. Success and failure analysis

As an effort to understand individual contributions to and/or failures of the place name features in performance increases, we analyzed their roles in some events. Biggest improvements were possible when the place information in the target events was clearly described, particularly with the hierarchical information. For instance, the improvement was more than 30% with the event containing “경상남도 김해시 지내동 동원아파트 뒤편 동대산”, which includes the province, city, town, apartment, and the name of the mountain behind the apartment, as the place name.

However, the event tracking task for some events was negatively influenced by our method as in Table 4. The first two below are the problems caused by place name ambiguities, and the third one is caused by the difficulty of even extracting place information:

- Similar events occurred in the same place
  When two distinct events occurred in the same place, two articles describing them respectively should be separated and fallen into two different event categories. However, Place information makes two separate specific incidents look similar if they are sub-events of a large event like Olympic Games, because the place where the games took place is the same. As a result, a variety of news articles on Olympic Games would be incorrectly classified into a very specific event when the place names are given too big a weight.

- The place name not specified or fixed
  Although two articles mention the same event, the place name found in the text may be too generic like “北한군 서해 도발 사건” (North Korean Navy’s attack in the West Sea). The place name West Sea is too general, and at the same time, the region could be referred to in other articles with different names like NIT or Northern Boundary.

- Not easy to extract a place name.

When the place name is part of the event name, it is difficult to extract it, because the place name is concatenated with other words without any spacing. Unless such a complex noun is decomposed into simple nouns, the current named entity recognizer will never find the place name. For instance, the word “간” (Cannes) in the event title “간영화제 수상” (awards in Cannes Film Festival) is the place where the film festival was held, but since the first four syllables are considered a single word, it is not possible to extract the place name unless noun segmentation is done first.

6. Evaluation with English news articles

For generalizability and extensibility of the proposed approach, we ran experiments using English corpus. The other reason is to put the experimental results using the Korean corpus in perspectives by using the de facto standard test collection that has been used in TDT research.
We used the multi-lingual test collection, TDT2 (Cieri, Graff, Liberman, Martey, & Strassel, 1999) provided by LDC, which consists of news articles from nine different sources including newswire, radio broadcasting, and TV news. Only six sources containing English documents were used for our experiments. Out of 64,527 English documents, we chose 53,620 eliminating those that are not news articles (e.g., TV entertainment programs). Table 5 shows some statistics of the test documents used for our experiments.

There are 100 topics in the collection: the first 37 topics are for training, the next 29 for system validation, and the rest (34) for testing. Fig. 4 shows a sample topic with title, definition, place, time, and related stories. As can be seen in Fig. 4, either the definition or the related stories can be used for topic tracking. In our work, we chose to use the topic definition part of 34 topics set aside for testing in the original collection. The number of relevant documents is 2161.

In a TDT system, a target story can be correctly detected (labeled as “Correct(+)” below), or missed (“Miss”). On the other hand, a non-target story can be left or undetected correctly (“Correct(-)”) or detected incorrectly (“False Alarm”). The relationship of the four different categories is shown in Fig. 5.

In order to make comparisons with previous research, especially that in Juha et al. (2004), we chose to use modified precision and recall, named $P_1$ and $R_1$ below, as well, in addition to traditional precision and recall. They evaluate the error rates based on false-alarm and miss with the following formulas (Juha et al., 2004):

$$P_1 = \frac{\#\text{correct}(+)}{\#\text{correct}(+) + \#\text{false alarm}}$$

$$R_1 = \frac{\#\text{correct}(+)}{\#\text{correct}(+) + \#\text{miss}}$$

$$F_1 = \frac{2 \times P_1 \times R_1}{P_1 + R_1}$$

6.1. Baseline evaluation

Unless baseline performance is at a reasonable level, any improvement over the baseline acquired with a new method would be meaningless. In order to ensure that our baseline system performance is comparable

---

Table 5
Data used for experiments

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th># News</th>
<th>Avg. size (byte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York Times (NYT)</td>
<td>Newswire</td>
<td>11,795</td>
<td>5280</td>
</tr>
<tr>
<td>Associated Press (APW)</td>
<td>Newswire</td>
<td>12,760</td>
<td>2134</td>
</tr>
<tr>
<td>CNN</td>
<td>Television</td>
<td>15,785</td>
<td>713</td>
</tr>
<tr>
<td>ABC</td>
<td>Television</td>
<td>2153</td>
<td>1304</td>
</tr>
<tr>
<td>Public Radio Inter. (PRI)</td>
<td>Radio</td>
<td>2913</td>
<td>1877</td>
</tr>
<tr>
<td>Voice of America (VOA)</td>
<td>Radio</td>
<td>8214</td>
<td>1322</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>53,620</td>
<td>2236</td>
</tr>
</tbody>
</table>

---

Fig. 4. A sample topic.

---

I. Asian Economic Crisis
- WHAT: The economic crisis in Asia
- WHERE: East and Southeast Asia
- WHEN: Beginning Fall 1997 and ongoing
- TOPIC EXPLICATION: Topic-related stories include: IMF bail-out and U.S. involvement, ripple effect in U.S. and other stock markets, impact on international business, impact on citizens of Asian countries.

Seed Article: CNN19980112.2130.0084

---

5 http://www.ldc.upenn.edu/.
with that of other state-of-art systems, we evaluated our system in terms of $P_1$ and $R_1$ to compare against (Juha et al., 2004), in addition to the traditional precision and recall. The comparison with the approach described in Juha et al. (2004) is particularly meaningful because it uses both time and place information and its performance was tested with TDT2. As in Table 6, we confirmed that our baseline performance is solid enough to proceed to test the proposed method using place information.

6.2. Evaluation of place information

The goals of the experiments using the English collection are basically the same as those with the Korean collection. We first examined the three different aspects (hierarchy, position, and coherence) of place information and then attempted to measure the improvement from explicitly using place information. The only difference in the experimental setup is that the case E was not considered because the experimental design in Juha et al. (2004) was different. We discuss a qualitative comparison with this approach in a later section.

Following are all the cases compared:

- Baseline: event tracking based on topical content only,
- Case A: baseline + use of hierarchical information of place names,
- Case B: baseline + position of place names,
- Case C: baseline + coherence of place names,
- Case D: A + B + C.

As can be seen in Table 7, use of place information enhances tracking effectiveness quite significantly, as much as 14.5% in Case D where all the features are used together. One reason for the larger magnitude of improvement than that of the experiment with the Korean corpus is that English corpus contains many international events that can be identified with a mention of the country name.

In order to see the importance of place information relative to time information for tracking, we ran further experiments with the English corpus. The results in Table 8 show that the maximum performance (74.07) was
obtained when both time and place information was used. Although the magnitude of the improvement over the use of publication dates (69.03 vs. 73.04) was meaningful, the contribution made by either place information was insignificant when both publication dates and extracted time information were used (73.87 vs. 74.07). Put differently, the time information was not helpful when place information was used. Our analysis shows that for “long-term” events, the time information was not specific enough to be helpful (e.g., Spring 1988) although additional time information was useful with a shorter-term events. Even for “short-term” events, a significant portion of them have less than five related documents which can be tracked by topicality only, making it unnecessary to use either place or time information.

Finally, we attempted to compare the performance of our approach with that of the previous one (Juha et al., 2004) where both time and place information is also used. As noted above, the evaluation was done with error-rate based precision and recall, named $P_1$ and $R_1$. Three different cases were compared:

Case A: the approach in Juha et al. (2004),
Case B: our approach with place information,
Case C: our approach with both time and place information.

As in Table 9, use of place information alone gave a marginal improvement over the previous approach, but when it is used in conjunction with time information, the performance difference was greater. The reason why the $P_1$ value in Case C is much smaller than that of Case 1 seems to stem from the fact that there was an extra training effort in Juha et al. (2004) by using seed on-target documents in addition to the topic description. In other words, there were more training documents for (Juha et al., 2004) than our system.

7. Conclusion

We addressed the issues related to using place names extracted automatically from news paper articles in the context of event tracking. Instead of extracting all place names from news articles and treat them as elements of vectors, we attempted to use a place name hierarchy for partial matching and characteristics of place name occurrences such as their relative locations and coherence in a news article.

As an effort to measure the impact of using place information separately from time information used in our recent work as well as other content-revealing vocabulary, we ran experiments using both Korean and English news articles. The results show that place information extracted from text and used separately from other content indeed help effectiveness of tracking in both languages. In English, in particular, place information seems to be more important than time information because a large proportion of news articles report on long-lasting events where time information is not very critical in identifying an event.
Our limited qualitative analysis indicates that relative contributions of time and place information vary depending on the nature of the events and their descriptions. The question of how to mix time and place information in computing similarity values depending on the nature of the events is the future work.

References


Smith, D. A. (2002). Detecting events with date and place information in unstructured text. In *JCDL’02* Portland, Oregon, USA.
