Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 327–343
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Cross-document event clustering using knowledge mining from co-reference chains

June-Jei Kuo, Hsin-Hsi Chen *

Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

Received 16 May 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

Unifying terminology usages which captures more term semantics is useful for event clustering. This paper proposes a metric of normalized chain edit distance to mine, incrementally, controlled vocabulary from cross-document co-reference chains. Controlled vocabulary is employed to unify terms among different co-reference chains. A novel threshold model that incorporates both time decay function and spanning window uses the controlled vocabulary for event clustering on streaming news. Under correct co-reference chains, the proposed system has a 15.97% performance increase compared to the baseline system, and a 5.93% performance increase compared to the system without introducing controlled vocabulary. Furthermore, a Chinese co-reference resolution system with a chain filtering mechanism is used to experiment on the robustness of the proposed event clustering system. The clustering system using noisy co-reference chains still achieves a 10.55% performance increase compared to the baseline system. The above shows that our approach is promising.

Keywords: Controlled vocabulary; Co-reference chains; Event clustering; Multi-document summarization

1. Introduction

News through the Internet is an important information source, is reported anytime and anywhere, and is disseminated across geographic barriers. Detecting the start of new events and tracking their progress (Allan, Carbonell, & Yamron, 2002; Chen & Ku, 2002; Chieu & Lee, 2004) are useful for decision-making in today's fast-changing network era. The research issues behind event clustering include: how many features are used to determine event clusters, which cue patterns are employed to relate news stories in the same event, how clustering strategies affect clustering performance using retrospective data or on-line data, how the time factor affects clustering performance, and how cross-document co-references are resolved.

Several studies, for example, text classification (Kolcz, Prabakarmurthi, & Kalita, 2001) and web-page classification (Shen, Chen, Yang, Zhang, & Lu, 2004), suggest that even simple summaries are quite effective in

* Corresponding author. Tel.: +886 2 33664888x311; fax: +886 2 23628167.
E-mail addresses: jjkuo@nlg.csie.ntu.edu.tw (J.-J. Kuo), hhchen@csie.ntu.edu.tw (H.-H. Chen).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.07.016

carrying over relevant information about a document. They showed that if a full-text classification method is directly applied to those documents, it incurs much bias for the classification algorithm, potentially losing focus on the main topic and important content. Moreover, for deeper document understanding, the co-reference chains (Cardie & Wagstaff, 1999) of documents capture information on co-referring expressions, i.e., all mentions of a given entity. Since the co-reference provides important clues to find text fragments containing salient information, various practical tasks can be done more reliably, i.e., text summarization (Azzam, Humphreys, & Gaizauskas, 1999; Chen, Kuo, Huang, Lin, & Wung, 2003), question answering (Lin, Chen, Liu, Tsai, & Wung, 2001; Morton, 1999), event clustering (Kuo & Chen, 2004), etc. In contrast, while producing summaries from multiple documents, cross-document co-reference analyses (Bagga & Baldwin, 1998; Gooi & Allan, 2004) continue their consideration if there are the same mentions of a name in different documents.

This paper shows that using summarization as pre-processing in event clustering is a viable and effective technique. Furthermore, we integrate co-reference chains from more than one document by unifying cross-document co-references of nominal elements. Instead of using the traditional clustering approaches, we propose a novel threshold model that incorporates time decay function and spanning window to deal with on-line streaming news. The rest of the paper is organized as follows. Section 2 reviews the previous work and shows our architecture. Section 3 describes a document summarization algorithm using co-reference chains. Section 4 tackles the issues surrounding mining controlled vocabulary. A normalized chain edit distance and two algorithms are proposed to incrementally mine controlled vocabulary from cross-document co-reference chains. Section 5 proposes an algorithm for on-line event clustering using dynamic threshold model. Section 6 specifies the data set and the experimental results, using the metric adopted by Topic Detection and Tracking (Fiscus & Doddington, 2002). A Chinese co-reference resolution system is introduced in Section 7, a chain filtering algorithm is proposed to improve the quality of auto-tagged co-reference chains and the related experimental results are shown. Finally, Section 8 is a conclusion.

2. Basic architecture

Kuo and Chen (2004) employed co-reference chains to cluster streaming news into event clusters. They think the co-reference chains and event words are complementary in some ways, hence they also introduced the event words as defined by Fukumoto and Suzuki (2000). Kuo and Chen's (2004) experimental results showed that both factors are useful. Furthermore, they present two approaches to combine the two factors for event clustering, which are called summation model and two-level model. The summation model simply adds the scores for both co-reference chains and event words. On the contrary, a two-level model is designed in such a way that the co-reference chains or the event words are used separately rather than simultaneously. However, the best performance was by the summation model and improved only 2%, in terms of detection cost, compared to the baseline system. One of the reasons is that the nominal elements used in cross-document co-reference chains may be different. The goal of this paper is to mine, incrementally, controlled vocabulary from co-reference chains of different documents for event clustering on streaming news.

Fig. 1 shows the architecture of event clustering. We receive documents from multiple Internet sources, such as newspaper sites, and then send them for document pre-processing. The pre-processing module deals with the sentence extraction and language idiosyncracy, e.g., Chinese segmentation and co-reference resolution. Document Summarization module analyzes each document and employs the co-reference chains and the related feature words, such as event words or high TF-IDF words, to produce the respective summaries. The controlled vocabulary mining module integrates the co-reference chains to generate controlled vocabulary automatically. Finally, the event clustering module uses weights of word features, and a similarity function to cluster the documents.

3. Document summarization using co-reference chains

Kuo and Chen (2004) only used the event words as features for clustering. The basic hypothesis is that an event word associated with a news article appears across in a number of paragraphs, but a topic word does not. Moreover, the domain dependency among words is a key clue to distinguish a topic and an event. This can be captured by *dispersion value* and *deviation value* (Fukumoto & Suzuki, 2000). The former tells if a word

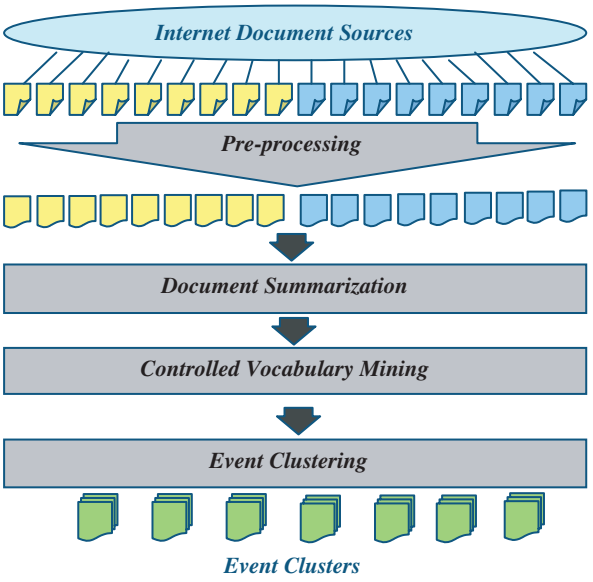

Fig. 1. Architecture of event clustering for streaming news.

appears across paragraphs (documents), and the latter tells if a word appears frequently. Event words are extracted by using these two values. Take an event "Air Accident of China Airlines" which happened on May 25, 2002 as an example. Each related news article has different event words, e.g., "body recovery", "set up", "17:10PM", "CKS airport", "Commander Lin", "stock market", "fly right negotiation", etc. Extracting such keywords is useful to understand the events, and distinguish one document from another. Nevertheless, due to the strict decision thresholds in the related formulas there are only a few event words extracted and we may lose some important feature words.

Thus, this paper introduces the higher TF-IDF words to be our document features. Document summarization module extracts the event words and the 20 highest TF-IDF words from each document. Then, the score of each sentence in a document is computed by adding three scores, i.e., the number of event words, the number of the highest TF-IDF words, and the co-reference scores discussed in the following paragraphs. Rather than using fixed number of sentences to generate summary, the sentence selection procedure is repeated until a dynamic number of sentences is retrieved. This number is equal to the compression rate multiplied by the total sentences in a document. For example, let the compression rate and total sentences be 0.35 and 15, respectively. In this case, the length of summary is 5, i.e., $0.35 * 15$.

The co-reference score of a sentence is computed as follows. The headlines of a news story can be regarded as its short summary. That is, in some sense, the words in the headline represent the content of a document. The co-reference chains that are initiated by the words in the headlines are assumed to have higher weights. A sentence which contains any words in a given co-reference chain is said to "cover" that chain. Those sentences which cover more co-reference chains contain more information, and should be selected to represent a document. Five scores shown below are computed sequentially to break the ties during sentence selection.

(1) For each sentence that is not selected, count the number of noun co-reference chains from the headline which are covered by this sentence and have not been covered by the previously selected sentences.
(2) For each sentence that is not selected, count the number of noun co-reference chains from the headline which are covered by this sentence, and add the count to the number of verbal terms in this sentence which also appear in the headline.
(3) For each sentence that is not selected, count the total number of noun co-reference chains which are covered by this sentence and have not been covered by the previously selected sentences.[1]

[1] This includes all co-reference chains and is not limited to co-reference chains from the headline.

(4) For each sentence that is not selected, count the total number of noun co-reference chains which are covered by this sentence and add the count to the number of verbal terms in this sentence which also appear in the headline.
(5) The position of a sentence.

Score 1 only considers nominal features. Score 2 considers both nominal and verbal features and both scores are initiated by the headlines. Scores 3 and 4 consider all the co-reference chains no matter whether these chains are initiated by the headline or not. These two scores rank those sentences on which scores 1 and 2 are tied. In addition, they can assign scores to news stories without headlines. Scores 1 and 3 are recomputed in the iteration. Finally, since news stories tend to contain more information in the leading paragraphs, Score 5 determines which sentence will be selected according to the position of a sentence, when sentences are of the same scores (1)–(4). The smaller the position number of a sentence, the more it is preferred.

4. Creating controlled vocabulary from individual co-reference chains

Streaming news stories are disseminated from different sources and written by different conventions and styles. The expression of an entity in a document may be different from the expression of the same entity in another document. Fig. 2 shows an example of four short co-reference chains in four different documents DOC1-DOC4, which are selected from our manual co-reference corpus.

Considering the co-reference chain in DOC1, "總統喬治．布希" (President George W. Bush) and "布希總統" (President Bush) denote the same person. There are two identical words "總統" (President) and "布希總統" (President Bush) between the chains in DOC1 and DOC2, so that word matching tells us these two chains have the same denotation. However, direct word matching between two co-reference chains may suffer from the following two problems.

(1) Because streaming news stories are disseminated from different sources anytime, the arrival sequence of documents affects the quality of controlled vocabulary. For example, when DOC3 arrives before DOC2, the two chains in DOC1 and DOC3 will denote two different named entities due to no word matching between the two co-reference chains. In this case, a resolution algorithm may miss some correct cross-document co-references.
(2) Because there are two matching words "總統" (President) and "他" (he) between the co-reference chains DOC2 and DOC4, they may be mistaken as the same person in spite of different person entities, i.e., "布希總統" (President Bush) and "總統柯林頓" (President Clinton). In this case, a resolution algorithm may produce incorrect cross-document controlled vocabulary.

4.1. Normalized chain edit distance

Instead of using word matching, the concept of normalized chain edit distance is proposed. The edit distance of two strings, s1 and s2, is defined as the minimum number of edit operations, i.e., insertions, deletions and substitutions, needed to transform string s1 to s2. Consider an example. Let strings s1 and s2 be defined as

DOC1: 總統(President) → 總統喬治．布希(President George W. Bush) → 布希總統(President Bush) → 小布希(Bush junior)
DOC2: 總統(President) → 布希總統(President Bush) → 他(he) → 他(he) → 布希(Bush junior) → 布希總統(President Bush) → 布希總統(President Bush) → 布希(Bush junior)
DOC3: 總統布希(President Bush) → 布希(Bush junior)
DOC4: 總統柯林頓(President Clinton) → 總統(President) → 他(he) → 柯林頓總統(President Clinton)

Fig. 2. Sample co-reference chains.

AAABB and BBAAA, respectively. The edit distance between s1 and s2 is calculated by the function edit_distance(s1, s2) and is 4. The smaller the edit distance, the more similar are the two strings. Here, the edit distance is extended to determine whether two given co-reference chains are similar or not. Assume there are two co-reference chains – say, *Given* and *Incoming*. Algorithm 1 computes the chain edit distance of *Incoming* and *Given* co-reference chains. If the score is smaller than a predefined threshold, the *Incoming* co-reference chain denotes the same entity as the *Given* co-reference chain, and is merged into *Given* chain in Algorithm 1. Otherwise, they are regarded as different entities.

Algorithm 1. Compute the normalized chain edit distance of incoming and given co-reference chains

1. Let len1 and len2 be the length (i.e., number of words) of *Incoming* and *Given* co-reference chains, respectively.
2. Let word1[i] and word2[j] be the *i*th and the *j*th elements in *Incoming* and *Given* co-reference chains, respectively.
3. Initialize score to be 0.
4. for $i = 1$ to len1 {
   min $= \infty$;
   $j = 1$ to len2 {
   (1) let $d =$ edit_distance(word1[i], word2[j])
   (2) $d = d$/max(length(word1[i]), length(word2[j]))
   (3) if $d <$ min then min $= d$}
   score + =min;}
5. Compute score = score/len1 and output the score.

Consider the sample co-reference chains shown in Fig. 2. Assume DOC1 and DOC2 are *Given* and *Incoming* co-reference chains, respectively. Let the threshold value, which is trained from 150 different co-reference chains, be 0.45. The normalized chain edit distance between these two co-reference chains is $(0 + 0 + 1 + 1 + 1/3 + 0 + 0 + 1/3)/8 = 0.33$. Hence, the two co-reference chains in DOC1 and DOC2 are deemed the same entity. Similarly, the edit distance between the *Given* chain in DOC1 and the *Incoming* chain in DOC4 is $(3/5 + 0 + 1 + 3/5)/4 = 0.55$. In contrast, this co-reference chain in DOC4 denotes a different entity from that in DOC1. On the other hand, although there is no matching word between the chains in DOC1 and DOC3, their normalized chain edit distance is low enough, i.e., $(3/7 + 1/3)/2 = 0.38$ (<0.45). Thus, these two chains can also be deemed to denote the same entity. In summary, the above two issues can be solved in Algorithm 1.

Pronouns (e.g., "他" (he)) and personal title words (e.g., "總統" (President)) are less specific in a co-reference chain, so that they contribute less information and are prone to incur errors in creating cross-document controlled vocabulary. DOC2 and DOC4 show an example. $((3/5 + 0 + 0 + 3/5)/4 = 0.30 < 0.45)$ which is an incorrect instance. In such a case, an alternative solution may be: pronouns and personal title words are excluded from cross-document co-reference chains during mining controlled vocabulary.

4.2. Creating controlled vocabulary

As temporal reference denotes a specific time or date, it is not meaningful to unify cross-document temporal references into controlled vocabulary. Thus, we ignore the temporal references in our approach. Algorithm 2 specifies how to mine controlled vocabulary incrementally. Fig. 3 shows some examples in controlled vocabulary. The term in bold font is a header (canonic form) of a unified co-reference chain.

Algorithm 2. Mining controlled vocabulary

1. Set the threshold value to be α.
2. Get the first news document and the accompanying co-reference chains.

(a) **中華航空公司**(China Airlines), 華航, 中華航空, 華航公司, China Airlines, China Air-lines, China-Airlines, 台灣最大的航空公司
(b) **澎湖**(Peng-Hu), 澎, 澎湖縣, 菊島
(c) **澎湖外海**(Sea around Peng-Hu), 外海, 澎湖馬公外海, 失事地點, 飛機失事海域, 出事現場
(d) **行政院**(The Executive Yan), 政院, 中央
(e) **交通部**(The Ministry of Transportation and Communications), 交部

Fig. 3. An example of controlled vocabulary.

3. Initialize the controlled vocabulary to be the co-reference chains.
4. Get the next news document and its co-reference chains until all are processed.
   a. Check each co-reference chain in the document sequentially.
   b. Check whether the co-reference chain is temporal expression or not. If yes, go back to step 4a. Otherwise, continue.
   c. Employ Algorithm 1 to compute the normalized chain edit distance between the co-reference chain (called *Incoming* co-reference chain) and each chain (called *Given* co-reference chain) in controlled vocabulary.
   d. If there is a normalized chain edit distance whose score is lower than α, the elements in *Incoming* chain are merged into the corresponding *Given* chain in the controlled vocabulary.
   e. If all the scores are larger than α, *Incoming* chain is new and a new entry is created in the controlled vocabulary. The longest term is regarded as a header, which will be used in clustering to unify the term usages.

4.3. Evaluation

We adopted the B-CUBED metric (Bagga & Baldwin, 1998) shown below to measure the precision and recall of the created controlled vocabulary.

$$\text{Precision}_i = \frac{\text{number of correct elements in the output chain containing entity } i}{\text{number of elements in the output chain containing entity } i} \quad (1)$$

$$\text{Recall}_i = \frac{\text{number of correct elements in the output chain containing entity } i}{\text{number of elements in the true chain containing entity } i} \quad (2)$$

The numerator of both formulas (1) and (2) means the number of the same elements between the true chain and the output chain for entity i. The final precision and recall rates are the average precision and recall rates of all entities. In addition, the F score uses the harmonic means of precision and recall. Besides direct evaluation, to evaluate the performance indirectly we also employed the created controlled vocabulary to the event clustering system proposed in Section 5.

4.3.1. Data set

In our experiment, we used the knowledge base provided by the United Daily News (http://udndata.com/), which has collected 6,270,000 Chinese news articles from six Taiwan local newspaper companies since 1975/1/1. To prepare a test corpus, we first set the topic to be "華航空難" (Air Accident of China Airlines), and the range of searching date from 2002/5/26 to 2002/9/4 when all rescue activities stopped. Total 964 related news articles, which each has a published date, news source, headline and content, are returned from search engine. All are in SGML format. After reading these news articles, we deleted five news articles which have headlines without any content. The average length of a news article is 15.6 sentences. All the above articles have also been manually tagged with co-reference chains. Furthermore, we asked three research assistants separately to merge the related co-reference chains into controlled vocabulary and then we used majority rule to create the gold answer.

4.3.2. Experimental results

Pronouns and personal title words occur frequently in co-reference chains. To verify if they have significant discrimination among chains, two alternative experiments are conducted. M1 used the original co-reference chains to create controlled vocabulary. In contrast, M2 excludes pronouns and personal title words in co-reference chains. The related F-scores are shown in Fig. 4. The threshold is α value in Algorithm 2. The baseline system only uses the word matching. Normalized chain edit distance is superior to word matching no matter which co-reference chains are adopted. The experimental results also verify that pronouns and personal title words in a co-reference chain contribute little information no matter whether word matching or edit distance approaches are employed. When the approach of edit distance using M2 with threshold 0.33 is adopted, the best performance, i.e., precision 96.49%, recall 96.67%, and F-score 96.58%, is achieved.

Analyzing the created controlled vocabulary using M2, we found that there are three major types of errors shown below.

(1) Ambiguous abbreviation problem, e.g., "澳門" (Macau) and "澳洲" (Australia) have the same abbreviation (i.e., "澳"), so that they were merged incorrectly.
(2) Lack of semantic information, e.g., "南部地區" (southern area) and "東部地區" (eastern area) were merged incorrectly.
(3) Word order problems, e.g., "雷馬遜颱風" (Remason Typhoon) cannot be merged with chain "颱風雷馬遜" (Typhoon Remason).

5. Event clustering

A single-pass complete link clustering algorithm incrementally divides the documents into several event clusters. Initially, the first document d_1 is assigned to cluster t_1, and the co-reference chains of d_1 form initial controlled vocabulary (refer to Steps 2–3 of Algorithm 2). Assume there already are k clusters when a new article d_i is considered. That is, clusters $t_1, t_2, \ldots, t_k$ $(k < i)$ have been detected. Document d_i may belong to one of k clusters, or it may form a new cluster t_{k+1}. This is determined by the similarity measure defined below.

At first, we mine new controlled vocabulary from current controlled vocabulary and the incoming news story. The procedure refers to Step 4 of Algorithm 2. Then we compute the similarities of the summary of the incoming news story with each summary in a cluster. The newly mined controlled vocabulary is global to each similarity computation. Let V_1 and V_2 be the vectors for the two summaries extracted from documents D_1 and D_2. Event clustering module uses the headers of the mined controlled vocabulary to replace the related terms in the processing summary. The 20 highest TF-IDF words are used as the feature words for each document. Moreover, whenever new documents are processed, the related feature words are recomputed. Each document is represented as a vector of normalized TF-IDF weights shown as follows.

$$w_{ij} = \frac{tf_{ij} \times \log \frac{N}{df_j}}{\sqrt{s_{i1}^2 + s_{i2}^2 + \cdots + s_{in}^2}} \tag{3}$$

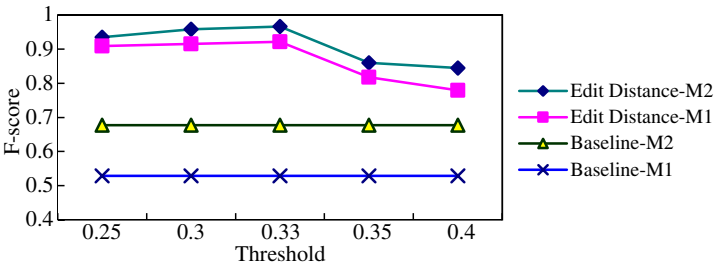

Fig. 4. *F*-measure using B-CUBED.

where tf_{ij} is frequency of term t_j in summary i, N is total number of summaries in the collection being examined, df_j is number of summaries that term t_j occurs, and s_{ij} denotes the TF-IDF value of term t_j in summary i.

The similarity between V_1 and V_2 is computed as follows.

$$\mathrm{Sim}(V_1, V_2) = \frac{\sum_{\text{common term } t_j} w_{1j} \times w_{2j}}{\sqrt{\sum_{j=1}^{n} w_{1j}^2}\sqrt{\sum_{j=1}^{m} w_{2j}^2}} \tag{4}$$

If the similarities of all document pairs are larger than a fixed threshold, the news story is assigned to the cluster. Otherwise, it forms a new cluster by itself.

The motivation for this approach is that news stories appearing on the stream closer together in time are more likely to contain discussion of the same event than those stories appearing further apart. Thus, instead of using a fixed detection threshold for comparison strategy, a dynamic detection threshold using a time decay function and spanning windows is proposed below (5). A dynamic detection threshold (d_th) is introduced, where th is an initial threshold. In other words, the earlier the documents are put in a cluster, the smaller their thresholds. Assume the publication day of document D_2 is later than that of document D_1.

$$d\_th(D_1, D_2) = \sqrt{\frac{\mathrm{dist}(D_1)/w\_\mathrm{size} + 1}{\mathrm{dist}(D_2)/w\_\mathrm{size} + 1}} \times th \tag{5}$$

where dist (denoted as day distance) denotes the number of days away from when the event happens, and $w_$size (denoted as window size) keeps the threshold unchanged within the same window.

6. Experimental results

6.1. Data set

The same data set described in Section 4.3.1 is used in this experiment. In addition, by referring the events classification of air accident used by Kuo and Chen (2004), the data set is classified as 13 events, e.g., rescue status. Meanwhile, two annotators are asked to read all the 959 news articles and classify these articles into 13 events or mark them as "other". A news article which reports more than one event may be classified into more than one event cluster. We then compare the classification results of annotators and only consider those results where the two annotators agree as our answer set. The distribution of the 13 focus events in the answer set are Fly right negotiation between Taiwan and Hong Kong (20), Cause of air accident (57), Confirmation of air accident (6), Influence on stock market (27), Influence on insurance fee (11), Influence on China Airlines (8), Influence on Peng-Hu archipelagoes (26), Punishment for persons in charge (10), News reporting (18), Wreckage found (28), Remains found (57), Rescue status (65), Solatium (34) and unused events (664) included "other" events and inconsistent coding. The number in the parentheses denotes the documents in the cluster.

6.2. Evaluation metric

We also adopt the metric used in topic detection and tracking (TDT) (Fiscus & Doddington, 2002). The evaluation is based on miss and false alarm rates and both rates are penalties. They can measure more accurately the behavior of users who try to retrieve news stories. If either rate is too high, users will not be satisfied with the clustering results. The performance is characterized by a detection cost, C_{Det}, in terms of the probability of miss and false alarm:

$$C_{\mathrm{Det}} = C_{\mathrm{Miss}} \times P_{\mathrm{Miss}} \times P_{\mathrm{target}} + C_{\mathrm{FA}} \times P_{\mathrm{FA}} \times P_{\text{non-target}} \tag{6}$$

where C_{Miss} and C_{FA} are costs of a miss and a false alarm, P_{Miss} and P_{FA} are the conditional probabilities of a miss and a false alarm, and P_{target} and $P_{\text{non-target}}(=1 - P_{\mathrm{target}})$ are the prior target probabilities. Manmatha, Feng, and Allan (2002) indicated that the standard TDT cost function used for all evaluations in TDT is

$C_{Det} = 0.02 \times P_{Miss} + 0.098 \times P_{FA}$. They think that false alarm should be penalized much more heavily than miss.

6.3. Experimental results

Table 1 shows the four model types used in the experiments. For comparison, the centroid-based single-pass clustering model is used as a baseline model. Conventional TF-IDF scheme selects 20 features for each incoming news article and each cluster uses 30 features as its centroid. Whenever an article is assigned to a cluster, the 30 words of the higher TF-IDFs are regarded as the new centroid of that cluster. In the co-reference model, the algorithm described in Section 3 studies the effects of document summarization using co-reference chains, and selects four sentences to represent the corresponding document. The experimental results with various thresholds are shown in Table 2. The best results of the two approaches are 0.012990 and 0.013137, respectively, when the threshold is set to 0.05.

With fixed thresholds strategies, the performance of co-reference model is worse than that of the centroid model. Hence, we study the effects of the dynamic thresholds described above in Section 5. Table 3 shows the results of co-reference model using various window sizes. The best detection cost, i.e., 0.012647, is achieved under window 2. Moreover, dynamic threshold using window size is more efficient than the best fixed threshold event clustering approach, i.e., $0.012647 < 0.012990$. For comparison, the best experimental results using Summation model (Kuo & Chen, 2004) are also shown in Table 3.

Next, we use the co-reference model to consider the length of each document's summary. Previous analysis was with fixed summary length. Dynamic lengths with different compression rates are now adopted. The detection cost using compression rate 0.35 is 0.011496, which is better than the costs of both the above co-reference

Table 1
Description of the four model types

Model type	Co-reference chain	Controlled vocabulary	Event words	Threshold type	TF-IDF sources	Removing pronouns and titles
Centroid	No	No	No	Fixed	Original document	No
Co-reference	Yes	No	No	Fixed	Fixed summary	No
Summation	Yes	No	Yes	Dynamic	Fixed summary	No
Final Model	Yes	Yes	Yes	Dynamic	Dynamic summary	Yes

Table 2
Detection cost for event clustering with fixed thresholds

Fixed threshold	Centroid model	Co-reference model
0.01	0.024644	0.015960
0.05	**0.012990**	**0.013137**
0.10	0.013736	0.015309
0.15	0.014331	0.016507
0.20	0.015480	0.016736
0.30	0.015962	0.017360

Table 3
Detection cost for event clustering with various window sizes (initial $th = 0.05$)

Window sizes	1	**2**	3	4
Co-reference model	0.012657	**0.012647**	0.012809	0.012942
Summation model	0.011223	**0.011603**	0.013109	0.013109

Table 4
Detection cost of co-reference model using dynamic length summary

Compression rate	0.25	0.3	**0.35**	0.4	0.45
Cdet	0.012074	0.012074	**0.011496**	0.011709	0.012181

Table 5
Detection cost of co-reference model using the proposed summarization module

Window size	1	**2**	3	4
Cdet	0.011828	**0.011083**	0.011817	0.011842

Table 6
Detection cost of a system with removing duplications

Threshold(α)	0.25	0.3	**0.33**	0.35
Cdet	0.011415	0.011407	**0.011407**	0.011554
Controlled vocabulary size	2023	1959	**1896**	1719

Table 7
Detection cost of a system without removing duplications

Threshold(α)	0.25	0.3	**0.33**	0.35
Cdet	0.011237	0.011183	**0.010966**	0.011084
Controlled vocabulary size	2143	2030	**2021**	1785

model using fixed length summary ($<$0.012647) and summarization model ($<$0.011603). The experimental results are shown in Table 4. We conclude that the flexible length summary conveys more information than the fixed length.

In addition to the summary length issue, we use the summarization module described in Section 3 to select sentences under different window sizes. Here, compression rate is 0.35 and threshold value is 0.04. Window size 2 further reduces detection cost to 0.011083 ($<$0.011946). The experimental results are shown in Table 5. It also shows the approach of including 20 highest TF-IDF words can select more informative sentences in document summarization.

Next, we introduce controlled vocabulary mined incrementally from co-reference chains. Tables 6 and 7 show the experimental results with and without removing duplications in the co-reference chains, respectively. We did not expect that the detection cost without duplication removal to be better than with removal. It seems that the more occurrences a word has in a co-reference chain, the more important it is.

In the final experiment, we kept the occurrences of topic elements except pronouns and personal title words, and mined controlled vocabulary from the resulting chains. As the quality of the controlled vocabulary is improved, the experimental results show that the performance of the final model is further improved to 0.010915. Comparing with the best detection costs of the centroid model (0.012990) and the summation model (0.011603), the best result of the final model has 15.97% and 5.93% performance gain.

7. Experiments using noisy co-reference chains

The experiments in Section 6 show that using either co-reference chains or controlled vocabulary improve the performance of the baseline system. Here we deal with the effects of noisy co-reference chains on cross-

document event clustering. In other words, the co-reference chains employed in the clustering are created automatically rather than manually. MUC (1998) indicated that the best F-measure of automatic co-reference resolution in English documents was 61.8%. To pinpoint the effects of controlled vocabulary in event clustering, we introduce a Chinese co-reference resolution system.

7.1. Flow of a Chinese co-reference resolution system

Fig. 5 shows the flow of a Chinese co-reference resolution system. The first four modules, including segmentation, named entity recognition (NER), part of speech tagging, and noun phrase chunking, aims to find the possible NP candidates. The statistical information for segmentation and tagging is extracted from Academia Sinica Balanced Corpus (ASBC, 1998). Then the attributes of the candidates are retrieved. Finally a co-reference resolution algorithm partitions the candidates into equivalence classes using the attributes.

Besides the named entities extracted by a Chinese NER system (Chen, Ding, Tsai, & Bian, 1998), we also employed NP chunkers to extract noun phrases. The maximal NPs, i.e., those NPs not covered by the other NPs, are selected as candidates.

We consider eight features to resolve Chinese co-reference relationships, including word/phrase itself, parts of speech of head nouns, named entity types, positions, number, pronouns, gender, and semantics of head nouns. To determine the feature values, some linguistic cues are employed. For example, we use morphemes such as "們" (men), "群" (qun), "對" (dui), and so on, to determine plurality. Monetary and percentage expressions are regarded as plural. Numerals are also a cue. Gender is determined by the cues proposed by Chen and Lee (1996). In Chinese, a married woman may place her husband's surname before her surname, and some Chinese characters have high score for male and some for female. The correct rate for gender assignment is 89%. In semantics part, we adopt Cilin senses (Mei, Zhu, Gao, & Yin, 1982), which is composed of 12 large categories, 94 middle categories, 1428 small categories, and 3925 word clusters. The sense tagging tool proposed by Chen, Lin, and Lin (2002) was adopted.

A clustering algorithm similar to Cardie and Wagstaff's (1999) is used to generate the co-reference chains. Using the same data set described in Section 4.3.1, the recall rate, the precision rate, and the F-measure of the co-reference resolution system are 57.52%, 34.28%, and 42.96%, respectively. The F-measure of co-reference resolution in Chinese documents is lower than that (61.8%) in English documents.

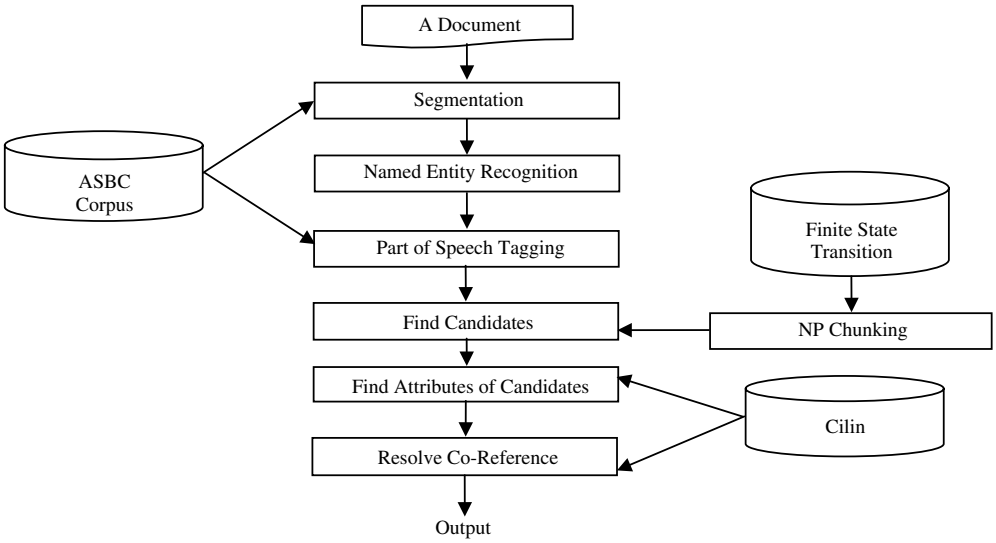

Fig. 5. Flow of a Chinese co-reference resolution system.

Table 8
Detection cost using noisy co-reference chains ($\alpha = 0.33$)

Initial threshold	0.01	**0.05**	0.10	0.15	0.20
Without controlled vocabulary	0.015988	**0.011809**	0.013878	0.015041	0.015712
With controlled vocabulary	0.014671	**0.012479**	0.014359	0.015252	0.016377

7.2. Experimental results of using noisy co-reference chains

Table 8 shows the results of using noisy co-reference chains. Compared with the cost of centroid model (*0.012990*), we still achieve a better detection cost (*0.011809*) in spite of the low quality of auto-tagged co-reference chains.

Because the metric of chain edit distance is adopted to generate the controlled vocabulary, the lower the quality of auto-tagged co-reference chains is, the worse the quality of controlled vocabulary. When we employ the controlled vocabulary mined from auto-tagged co-reference chains, the performance is worse than without using controlled vocabulary. The size of controlled vocabulary is 2230 and the best detection cost with controlled vocabulary is increased to 0.012479. There are two major types of errors:

(1) Noun phrase errors
Boundary errors during segmentation, NER and NP chunking may result in wrong NPs. The chain edit distance is increased for wrong NPs, so the boundary errors keep two co-reference chains denoting the same entities from being merged into the same controlled vocabulary.
(2) Accuracy of co-reference chains
The precision of an auto-tagged co-reference chain may incur wrong controlled vocabulary. Consider a wrong co-reference chain "中正機場 (CKS airport) → "關西機場 (Kansai airport) → "機場 (airport). Here, "關西機場 (Kansai airport) in Japan is wrongly deemed as the same airport "中正機場" (CKS airport) in Taiwan.

7.3. A co-reference chain filter

The co-reference resolution system specified in Section 7.2 employs only the information in a document to find the NP candidates and resolve co-references. Here we use a document set in a deferral period to revise a co-reference chain. The concept of deferral period is similar to topic detection and tracking (Chen & Ku, 2002; Allan et al., 2002), which defers the decision of story segmentation, new event detection, topic detection, and link detection to a period, e.g., the collection of 10 news stories. The deferral period keeps the incremental capability of Algorithm 2.

For each term in an auto-tagged co-reference chain, we extract all the sentences containing the term from a document set. For each extracted sentence, we include the previous and the following two sentences to form a context for this term. In other words, the sentence span is 5. In this way, we derive contexts for each term in a co-reference chain. Using the context, a term extraction algorithm similar to Chien (1997) is employed. The related noun phrases are corrected according to the newly extracted terms. Then we compare the similarity of contexts of each term pair in a chain. Those term pairs having similarity score above a threshold are kept in the chain.

Two approaches are adopted to measure the context similarity. The first one is an *Overlap Ratio* method defined as follows.

$$\text{OverlapRatio} = \frac{|\text{Context}(t_1) \cap \text{Context}(t_2)|}{\min(|\text{Context}(t_1)|, |\text{Context}(t_2)|)} \tag{7}$$

Table 9
Overlap ratios of some example term pairs

Item	No.					
	1		2		3	
Term Pair	澎湖	菊島	中華航空公司	國泰航空公司	中華航空公司	華航
Co-occurring terms	8866	282	860	60	860	17140
Same terms	239		17		746	
Overlap ratio	0.85 (239/283)		0.28 (17/60)		0.87 (746/860)	

where

Context(t_1) and Context(t_2) denote the contexts of terms t_1 and t_2, respectively,
Context(t_1) ∩ Context(t_2) denotes the term overlap, and
|Context(t_1)|, |Context(t_2)|, and |Context(t_1) ∩ Context(t_2)| denote total number of different terms in the designated contexts.

Some examples are shown in Table 9. The correct term pair has a larger ratio, in contrast, the wrong one has a smaller ratio.

The second approach is *Chi-square test* common used in collocation extraction (Manning & Schutze, 1999). The Chi-square test can be applied to any size table. Taking r-by-s table as an example, the following formula is used to compute the Chi-square value of two terms in a co-reference chain. Then, a probability level α and degree of freedom $(r-1 * s-1)$ are used to obtain the critical value by looking up the Chi-square table. Assume the null hypotheses H_0 and H_1 are the two terms denoting the same entities and different entities, respectively. If the Chi-square value is smaller than critical value, the null hypothesis is accepted. Otherwise, the null hypothesis is rejected and the term pair is removed from co-reference chain. An example using Chi-square test is shown in Appendix A.

$$\kappa^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(f_{ij} - \frac{f_{i.}f_{.j}}{N}\right)^2}{\frac{f_{i.}f_{.j}}{N}} \tag{8}$$

where

f_{ij} denotes the number in row i and column j,
$f_{i.}$ denotes the sum of numbers in row i,
$f_{.j}$ denotes the sum of numbers in column j, and
$N = \sum_{i=1}^{r} f_{i.} + \sum_{j=1}^{s} f_{.j}$ computes the total number.

For evaluating these two approaches, we randomly selected 2400 correct word pairs and 2400 wrong word pairs from the above noisy co-reference chains. Thus, if we just guess by chance in such a way that all are correct or all are wrong, the probable accuracy is 0.5, which can be deemed as the lower bound for accuracy. Table 10 lists the experimental results of these two approaches under different settings. To decrease the number of false positives in the co-reference chains and tolerate the false alarms in noisy co-reference chains, the level of significance of P is set to 0.05 for Chi-square testing rather than using a smaller value. Since the test corpus is collected from the streaming news for more than 3 months, we introduce *the frequency difference measure* (FD) to ignore the fluctuation of event focus. For example, the frequency of the word "搜救" (rescue) in the wreckage-found event is much larger than that in the cause of air accident event. Those terms which have high frequency difference easily created the noise in Chi-square testing and should be excluded from the observed words. In addition, for the Chi-square approximation to be valid, the observed frequency should be at least 5. Table 10(a) demonstrates the accuracy under various frequency differences. The best result 0.525625 is achieved in the case that frequency difference is equal to 30. Although all the accuracies are above 0.5, there

Table 10
Accuracy of co-reference chain filter

(a) Using Chi-square testing only (P = 0.05)					
Frequency difference	20	25	**30**	35	40
Chi-square only	0.516667	0.523333	**0.525625**	0.524583	0.518750
(b) Using overlap ratio testing only					
Ratio threshold	0.55	0.60	0.65	**0.70**	0.75
Ratio only	0.586250	0.597292	0.602708	**0.607692**	0.603333
(c) Using both overlap ratio and chi-square testing (P = 0.05)					
Ratio threshold	0.55	**0.60**	0.65	0.70	0.75
Ratio + Chi (FD = 40)	0.668333	0.671667	0.667917	0.661875	0.646667
Ratio + Chi (**FD = 35**)	0.668750	**0.671875**	0.668745	0.662917	0.647500
Ratio + Chi (FD = 30)	0.667292	0.670417	0.667708	0.663125	0.647917
Ratio + Chi (FD = 25)	0.665208	0.668750	0.667083	0.662708	0.648125
Ratio + Chi (FD = 20)	0.660625	0.665208	0.665417	0.661667	0.647083

Table 11
Detection cost using clearer co-reference chains (FD = 35, $\alpha = 0.33$)

Ratio threshold	0.55	**0.6**	0.65	0.7	0.75
Controlled vocabulary size	1434	**1392**	1336	1260	1123
Cdet (initial th = 0.05)	0.011793	**0.011620**	0.011908	0.012128	0.012137

is of no statistically significant difference. Table 10(b) shows that *Overlap Ratio* is more effective. *Overlap Ratio only* outperforms the *Chi-square testing only*. When combining these two approaches in Table 10(c) the accuracy is improved significantly (refer to threshold = 0.60 and FD = 35).

Consider an example to demonstrate why combining these two approaches is more effective. The frequencies of the terms "華航" ("China Airlines") and "台華輪" ("Tai-Hwa ship") are 17,140 and 357, respectively, and the number of their co-occurring terms is 269. If we use the *Overlap Ratio* measure only (269/357=0.76), the term pair is deemed to denote the same entity. When we further use the *Chi-square test* to verify this term pair, the corresponding *Chi-square value* is larger than *the critical value*, i.e., they are different terms. Thus, this term pair is not considered as the same entity and is excluded from the co-reference chain. In this way, false positives are decreased and the overall performance is improved.

7.4. Performance of event clustering using clearer co-reference chains

The co-reference chain filter not only revises the error terms in co-reference chains, but also deletes the wrong term pairs. In this way, the quality of auto-tagged co-reference chains is improved, so that the quality of the created controlled vocabulary is improved as well. Appendix B demonstrates an example of controlled vocabulary before/after employing chain filter.

Table 11 shows the performance of event clustering using clearer co-reference chains. Compared with the best result of the centroid model (Table 2, 0.012990), our performance is still better. We further compare the best cost (0.011620) from Table 11 using clearer co-reference chains with that (0.012479) using the noisy co-reference chains from Table 8. The performance is improved 6.88%. That indicates the co-reference chains are cleaned indirectly. Besides, the cost (0.011620) from Table 11 with controlled vocabulary is better than that (0.011809) without controlled vocabulary from Table 8. Thus, we conclude that controlled vocabulary is promising in event clustering regardless of using manual or auto-tagged co-reference chains.

8. Concluding remarks

This paper proposes a normalized chain edit distance to mine, incrementally, controlled vocabulary from cross-document co-reference chains, and uses the results to unify the features used in event clustering on

streaming news. Time decay function and spanning window capture the specific characteristics of on-line news. The experiments using manual co-reference chains show that occurrences of discriminative elements in a chain are useful, and pronouns as well as personal title words may introduce errors. The final model demonstrates 15.97% and 5.93% improvement compared to the centroid and the summation models, respectively. Furthermore, a Chinese co-reference resolution system is introduced to investigate the performance of event clustering under a noisy environment. A chain filtering algorithm is proposed, and the related experiments show the positive effects of using both co-reference chains and the controlled vocabulary in event clustering. In the future, we plan to apply controlled vocabulary to other applications, such as summary generation and construction of named entity ontology. Furthermore, we will also use the documents of other topics to study the validity of our proposed model in event clustering.

Acknowledgement

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC94-2752-E-001-001-PAE and NSC95-2752-E001-001-PAE.

Appendix A

An example of Chi-square test for a term pair "澎湖" (Peng-Hu) and "菊島" (The island of chrysanthemum, alias of Pen-Hu) using constraints of both frequency > =5 and frequency difference < =10.

Item	風情	海鱸	夏天	大力	少了	年夏	廣場	情人	藝人
澎湖	11	7	5	5	5	5	5	5	5
菊島	17	14	13	8	15	11	13	10	5
$f_{.j}$	28	21	18	13	20	16	18	15	10

Item	七夕	休閒	市長	的澎	夏季	藍色	菊島	日遊	今夏
澎湖	6	6	7	7	8	8	37	19	9
菊島	10	8	16	17	11	14	40	24	13
$f_{.j}$	16	14	23	24	19	22	77	43	22

Item	水族館	海鱸節	馬英九	千萬風情	三日遊	優惠折扣	$f_{i.}$
澎湖	10	5	5	5	9	9	203
菊島	12	13	14	6	9	9	322
$f_{.j}$	22	18	19	11	18	18	*525*

Assume H_0 is: "澎湖" (Peng-Hu) and "菊島" (The island of chrysanthemum, alias of Pen-Hu) denote the same entity, and H_1 is: two terms denote the different entity.

$$\kappa^2 = \sum_{i=1}^{2} \sum_{j=1}^{24} \frac{\left(f_{ij} - \frac{f_{i.}f_{.j}}{525}\right)^2}{\frac{f_{i.}f_{.j}}{525}} = 14.7234$$

A level of significance $P = 0.05$ is selected. The critical value of Chi-square with $(2-1)x(24-1)$ degree of freedom is ***35.17***. As the Chi-square score is smaller than the critical value (*14.7234 < 35.17*), we can conclude that "澎湖" (Peng-Hu) and "菊島" (The island of chrysanthemum) denote the same entity based on the hypothesis H_0.

Appendix B. Controlled vocabulary before/after employing chain filter

(Before using the chain filter)

(1.a) 1996年環球航空客機

六一一班機,飛機, 班機,六一一班次飛機, 機艙, 班飛機,驚人的飛機, 失事的班機, 該班機, 最好的飛機, 老舊飛機, 機身, 客機, 空難班機, 失事班機, 失事飛機, 距離飛機, 失事客機, 747型客機 ,727客機, 該架班機, 運輸機, 機率, 這架班機,每班飛機, 七三八型飛機, 頭等飛機, 機件 ,七四七型班機, 死亡班機, 民航客機, 機械,型貨機, 機體, 失事的班機機型, 這班死亡班機, 這班飛機, 這架失事班機, 指示飛機, 搭飛機, 1996年環球航空客機, 該機, 737客機空中, 前機, 這班失事的班機, 三點的班機, 一些飛機, 空難飛機, 這些飛機, 貨機, 這次飛機, 國際機場, 不明航機, 航機, 人員指示飛機, 新飛機, 不同飛機, 飛機機艙, 機腹, 任何異狀的飛機, 航空飛機, 小飛機, 這班失事客機, 主機身, 安全的飛機, 軍用運輸機, 老舊的客機, 其他班機, 交機, 這架飛機, 政府主管機關, 兩架飛機 ,空難事故飛機, 六一四班機, ,國內機場, 這架客機, 那架飛機, 靜止的飛機, 民航機, 小的飛機, 環球航空班機, 全部下機, 多小時的班機, 空難的死亡班機, 機翼, 北方航空空難班機, 機頭, 機尾, 這型飛機機身, 八００型班機, 其他飛機, 八００型客機, 該架飛機, 大機尾, 此時飛機, 同型機種, 二００型客機, 舊機, 該架失事班機, 部分的機體, 疑為飛機, 包機, 不同時段的班機, 港龍班機, 機票, 錄影機, 不少包機, 馬公機場, 半小時的飛機, 商務艙, 前駕駛艙, 座艙, 失事客機機, 路線的商用客機, 環球航空班機空中, 部分機翼, 二００型飛機, 其中座艙, 611班機座艙, 隱型機, ,四號發動機, 這時飛機, 駕駛艙, 四十公尺的飛機機艙, 北高航線機, 老飛機, 機種, 該款客機, 機型, 五十三架飛機, 搭包機, 七四七四百型客機, 尾翼

(1.b) 澎湖外海

澎湖外海, 乘客外, 外海, 紐約外海, 新竹外海, 港外海

(1.c) 失事的747-200型客機

機身, 747-200客機, 勤的班機, 海的747-200客機, 航班的747-200客機, 新的747-400客機, 飛機, 失事的747-200班機, 發動機 ,班機, 747-200型飛機, 型飛機, 客機, 失事的747-200型客機, 747-200型客機

(1.d) 熱帶性低氣壓環流影響

雙重影響, 低氣壓影響, 影響, 熱帶性低氣壓環流影響

(1.e) 短暫陣雨

陣雨, 雷雨, 短暫陣雨, 大雷雨

(After using the chain filter)

(2.a) 六一一班次飛機

六一一班機, 六一一班次飛機, 班機, 飛機

(2.b) 澎湖外海

澎湖外海, 紐約外海, 外海

(2.c) 失事的747-200型客機

客機, 747-200型客機, 失事的747-200型客機

(2.d) 低氣壓影響

雙重影響, 低氣壓影響, 影響

(2.e) 短暫陣雨

陣雨、雷雨、短暫陣雨

References

Allan, J., Carbonell, J., & Yamron, J. (Eds.). (2002). *Topic detection and tracking: Event-based information organization*. Massachusettes: Kluwer Academic Publisher.

ASBC (1998). Academia Sinica balanced Corpus. Technique report, CKIP, Academia Sinica, 98-03.

Azzam, S., Humphreys, K. & Gaizauskas, R. (1999). Using coreference chains for text summarization. In *Proceedings of 37th annual meeting of ACL workshop on coreference and its applications* (pp. 77–84). College Park, Maryland.

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th annual meeting of ACL and the 17th international conference on computational linguistics* (pp. 79–65). Montreal, Quebec.

Cardie, C. & Wagstaff, K. (1999). Noun phrase co-reference as clustering. In *Proceedings of the joint conference on EMNLP and VLC* (pp. 82–89). College Park, Maryland.

Chen, H. H., Ding, Y. W., Tsai, S. C. & Bian, G. W. (1998). Description of the NTU system used for MET2. In *Proceedings of 7th MUC*. Fairfax, <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html>.

Chen, H. H., & Ku, L. W. (2002). An NLP & IR approach to topic detection. In *Topic detection and tracking: Event-based information organization* (pp. 243–264). Massachusettes: Kluwer Academic Publisher.

Chen, H. H., Kuo, J. J., Huang, S. J., Lin, C. J., & Wung, H. C. (2003). A summarization system for Chinese news from multiple sources. *Journal of American Society for Information Science and Technology, 54*(13), 1224–1236.

Chen, H. H., & Lee, J. C. (1996). Identification and classification of proper nouns in Chinese texts. In *Proceedings of 16th COLING* (pp. 222–229). Copenhagen, Denmark.

Chen, H. H., Lin, C. C., & Lin, W. C. (2002). Building a Chinese–English WordNet for translingual applications. *ACM Transactions on Asian Language Information Processing, 1*(2), 103–122.

Chien, L. F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of the 20th ACM SIGIR conference* (pp. 50–58). Philadelphia, PA.

Chieu, H. L., & Lee, Y. K. (2004). Query based event extraction along a timeline. In *Proceedings of the 27th ACM SIGIR conference* (pp. 425–432). Sheffield, UK.

Fiscus, J. G., & Doddington, G. R. (2002). Topic detection and tracking evaluation overview. In *Topic detection and tracking: Event-based information organization* (pp. 17–32). Massachusettes: Kluwer Academic Publisher.

Fukumoto, F., & Suzuki, Y. (2000). Event tracking based on domain dependency. In *Proceedings of the 23rd ACM SIGIR 2000 conference* (pp. 57–64). Athens, Greece.

Gooi, C. H., & Allan, J. (2004). Cross-document coreference on a large corpus. In *Proceedings of human language technology conference/ North American chapter of the association for computational linguistics annual meeting* (pp. 9–16). Boston, MA.

Kolcz, A., Prabakarmurthi, V. & Kalita, J. (2001). Summarization as feature selection for text categorization. In *Proceedings of tenth international conference on information and knowledge management* (pp. 365–370). Atlanta, Georgia.

Kuo, J. J. & Chen, H. H. (2004). Event clustering on streaming news using co-reference chains and event words. In *Proceedings of the ACL workshop on coreference and its applications* (pp. 17–23). Barcelona, Spain.

Lin, C. J., Chen, H. H., Liu, C. C., Tsai, C. H. & Wung, H. C. (2001). Open domain question answering on heterogeneous data. In *Proceedings of ACL workshop on human language technology and knowledge management* (pp. 79–85). Toulouse, France.

Manmatha, R., Feng, A. & Allan, J. (2002). A critical examination of TDT's cost function. In *Proceedings of the 25th ACM SIGIR conference* (pp. 403–404). Tampere, Finland.

Manning, C. D., & Schutze, H. (1999). *Fundamental statistical natural language processing*. London: The MIT Press.

Mei, C. J., Zhu, Y., Gao, Y. C., & Yin, H. X. (1982). *tong2yi4ci2ci2lin2*. Shanghai: Shanghai Dictionary Press.

Morton, T. S. (1999). Using coreference for question answering. In *Proceedings of 37th annual meeting of ACL workshop on coreference and its applications* (pp. 85–89). College Park, Maryland.

MUC (1998). In *Proceedings of 7th message understanding conference*. Fairfax, VA,, <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html>.

Shen, D., Chen, Z., Yang, Q., Zeng H. J., Zhang, B. Y. & Lu, Y. C. (2004). Web-page classification through summarization. *Proceedings of the 27th ACM SIGIR conference* (pp. 242–249). Sheffield, UK.