Event Clustering on Streaming News Using Co-Reference Chains and Event Words

June-Jei Kuo
Department of Computer Science and
Information Engineering
National Taiwan University, Taipei, Taiwan
jjkuo@nlg.csie.ntu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and
Information Engineering
National Taiwan University, Taipei, Taiwan
hh_chen@csie.ntu.edu.tw

Abstract

Event clustering on streaming news aims to group documents by events automatically. This paper employs co-reference chains to extract the most representative sentences, and then uses them to select the most informative features for clustering. Due to the long span of events, a fixed threshold approach prohibits the latter documents to be clustered and thus decreases the performance. A dynamic threshold using time decay function and spanning window is proposed. Besides the noun phrases in co-reference chains, event words in each sentence are also introduced to improve the related performance. Two models are proposed. The experimental results show that both event words and co-reference chains are useful on event clustering.

1 Introduction

News, which is an important information source, is reported anytime and anywhere, and is disseminated across geographic barriers through Internet. Detecting the occurrences of new events and tracking the processes of the events (Allan, Carbonell, and Yamron, 2002) are useful for decision-making in this fast-changing network era. Event clustering automatically groups documents by events that are specified in the documents in a temporal order. The research issues behind event clustering include: how many features can be used to determine event clusters, which cue patterns can be employed to relate news stories in the same event, how the clustering strategies affect the clustering performance using retrospective data or on-line data, how the time factor affects clustering performance, and how multilingual data is clustered.

Chen and Ku (2002) considered named entities, other nouns and verbs as cue patterns to relate news stories describing the same event. A centroid-based approach with a two-threshold scheme determines relevance (irrelevance) between a news story and a topic cluster. A least-recently-used removal strategy models the time factor in such a way that older and unimportant terms will have no effect on clustering. Chen, Kuo and Su (2003) touched on event clustering in multilingual multi-document summarization. They showed that translation after clustering is better than translation before clustering, and translation deferred to sentence clustering, which reduces the propagation of translation errors, is most promising.

Fukumoto and Suzuki (2000) proposed concepts of topic words and event words for event tracking. They introduced more semantic approach for feature selection than the approach of parts of speech. Wong, Kuo and Chen (2001) employed these concepts to select informative words for headline generation, and to rank the extracted sentences in multi-document summarization (Kuo, Wong, Lin, and Chen, 2002).

Bagga and Baldwin (1998) proposed entity-based cross-document co-referencing which uses co-reference chains of each document to generate its summary and then use the summary rather than the whole article to select informative words to be the features of the document. Azzam, Humphreys, and Gaizauskas (1999) proposed a primitive model for text summarization using co-reference chains as well. Silber and McCoy (2002) proposed a text summarization model using lexical chains and showed that proper nouns and anaphora resolution is indispensable.

The two semantics-based feature selection approaches, i.e., co-reference chains and event words, are complementary in some sense. The former denotes equivalence classes of noun phrases, and the latter considers both nominal and verbal features, which appear across paragraphs. This paper will employ both co-reference chains and event words for temporal event clustering. An event clustering system using co-reference chains is described in Section 2. The evaluation method and the related experimental results are described in Section 3. The event words are introduced and discussed in Section 4. Section 5 proposes a summation model and a two-level model, respectively for event clustering using both co-reference chains and event words. Section 6 concludes the remarks.

2 Event Clustering using Co-Reference Chains

A co-reference chain in a document denotes an equivalence class of noun phrases. (Cardie and Wagstaff, 1999) A co-reference resolution procedure is first to find all the possible NP candidates. It includes word segmentation, named entity extraction, part of speech tagging, and noun phrase chunking. Then the candidates are partitioned into equivalence classes using the attributes such as word/phrase itself, parts of speech of head nouns, named entities, positions in a document, numbers (singular, plural, or unknown), pronouns, gender (female, male, or unknown), and semantics of head nouns. As the best F-measure of automatic co-reference resolution in English documents in MUC-7 was 61.8% (MUC, 1998), a corpus hand-tagged with named entities, and co-reference chains are prepared and employed to examine the real effects of co-reference chains in event clustering r.

Headlines of a news story can be regarded as its short summary. That is, the words in the headline represent the content of a document in some sense. 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 are selected to represent a document. Each sentence in a document is ranked according to the number of co-reference chains that it covers. Five scores shown below are computed. Sentences are sorted by the five scores in sequence and the sentences of the highest score are selected. The selection procedure is repeated until the designated number of sentences, e.g., 4 in this paper, is obtained.

(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 number of noun co-reference chains, which are covered by this sentence and have not been covered by the previously selected sentences.
(4) For each sentence that is not selected, count the 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. Comparatively, Score 2 considers both nominal and verbal features together. Both scores are initiated by headlines. Scores 3 and 4 consider all the co-reference chains no matter whether these chains are initiated by headlines or not. These two scores ranks those sentences of the same scores 1 and 2. Besides, 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 position of sentences, when sentences are of the same scores (1)-(4). The smaller the position number of a sentence is, the more it will be preferred.

The sentences extracted from a document form a summary for this document. It is in terms of a term vector with weights defined below. It is a normalized TF-IDF.

$$w_{ij} = \frac{tf_{ij} \times \log \frac{N}{df_j}}{\sqrt{s_{i1}^2 + s_{i2}^2 + \cdots + s_{in}^2}} \tag{1}$$

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.

A single-pass complete link clustering algorithm incrementally divides the documents into several event clusters. We compute the similarities of the summary of an incoming news story with each summary in a cluster. Let V_1 and V_2 be the vectors for the two summaries extracted from documents D_1 and D_2. The similarity between V_1 and V_2 is computed as follows.

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

If all the similarities are larger than a fixed threshold, the news story is assigned to the cluster. Otherwise, it forms a new cluster itself. Life span is a typical phenomenon for an event. It may be very long. Figure 1 shows the life span of an air crash event is more than 100 days. To tackle the long life span of an event, a dynamic threshold (d_th) shown below is introduced, where th is an initial threshold. In other words, the earlier the documents are put in a cluster, the smaller their thresholds are. Assume the published day of document D2 is later than that of document D1.

$$d\_th(D_1, D_2) = \sqrt{\frac{dist(D_1)/w\_size+1}{dist(D_2)/w\_size+1}} \times th \qquad (3)$$

where dist (day distance) denotes the number of days away from the day at which the event happens, and w_size (window size) keeps the threshold unchanged within the same window.

Moreover, we use square root function to prevent the dynamic threshold from downgrading too fast.

3 Test Collection

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 6 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 (stopping all rescue activities). Total 964 related news articles, which have published date, news source, headline and content, respectively, are returned from search engine. All are in SGML format. After reading those news articles, we deleted 5 news articles which have headlines but without any content. The average length of a news article is 15.6 sentences. Figure 1 depicts the distribution of the document number within the event life span, where the x-axis denotes the day from the start of the year. For example, "146" denotes the day of '2002/5/26', which is the 146th day of year 2002.

Then, we identify thirteen focus events, e.g., rescue status. Meanwhile, two annotators are asked to read all the 959 news articles and classify these articles into 13 events. If a news article can not be classified, the article is marked as "other" type. A news article which reports more than one event may be classified into more than one event cluster. We compare the classification results of annotators and consider those consistent results as our answer set. Table 1 shows the distribution of the 13 focus events.

Event Name	Number of Documents
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 Airline	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
Other	664

Table 1: Focus Events

We adopt the metric used in Topic Detection and Tracking (Fiscus and Doddington, 2002). The evaluation is based on miss and false alarm rates. Both miss and false alarm are penalties. They can

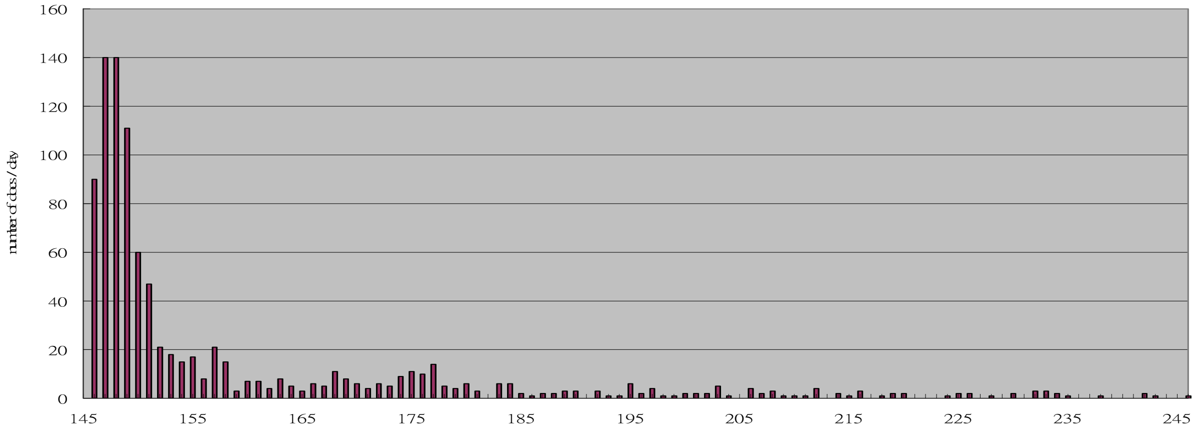

Figure 1. Event Evolution of China Airlines Air Accident (2002/5/26 ~ 2002/9/4)

measure more accurately the behavior of users who try to retrieve news stories. If miss or false alarm 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:

$$C_{Det} = C_{Miss} \times P_{Miss} \times P_{\mathrm{target}} + C_{FA} \times P_{FA} \times P_{non-t\mathrm{arget}} (4)$$

where C_{Miss} and C_{FA} are costs of a miss and a false alarm, respectively, P_{Miss} and P_{FA} are the conditional probabilities of a miss and a false alarm, and $P_{t\arg et}$ and $P_{non-t\arg et}\left(=1-P_{t\arg et}\right)$ are the prior target probabilities.

Manmatha, Feng and Allan (2002) indicated that the standard TDT cost function used for all evaluations in TDT is $C_{\det} = 0.02P_{Miss} + 0.098P_{FA}$, when C_{Miss}, C_{FA} and P_{target} are set to 1, 0.1 and 0.02, respectively. The less the detection cost is, the higher the performance is.

For comparison, the centroid-based approach and single pass clustering is regarded as a baseline model. Conventional TF-IDF scheme selects 20 features for each incoming news articles and each cluster uses 30 features to be 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. The experimental results with various thresholds are shown in Table 2. The best result is 0.012990 when the threshold is set to 0.05.

Fixed Threshold	$C_{\det}$
0.01	0.024644
0.05	**0.012990**
0.10	0.013736
0.15	0.014331
0.20	0.015480
0.25	0.015962

Table 2: Detection Costs Using Centroid Approach

Kuo, Wong, Lin and Chen (2002) indicated that near 26% of compression rate is suitable for a normal reader in multi-document summarization. Recall that the average length of a news story is 15.6 sentences. Following their postulation, total 4 sentences, i.e., 16/4, are selected using co-reference chains. Table 3 shows the detection cost with various threshold settings. We found that the best result could be obtained using threshold 0.05, however, it was lower than the result of baseline (i.e., 0.013137 > 0.012990).

Next, we study the effects of dynamic thresholds. Three dynamic threshold functions are experimented under the window size 1. A linear decay approach removes the square root function in Formula (3). A slow decay approach adds a constant (0.05) to Formula (3) to keep the minimum threshold to be 0.05 and degrades the threshold slowly. Table 4 shows that Formula (3) obtained the best result, and the dynamic threshold approach is better than the baseline model.

Fixed Threshold	$C_{\det}$
0.01	0.015960
0.05	**0.013137**
0.10	0.015309
0.15	0.016507
0.20	0.016736
0.25	0.017360

Table 3. Detection Costs Using Co-Reference Chains

Function Type	Linear decaying	Formula (3)	Slow Decaying
$C_{\det}$	0.013196	**0.012657**	0.016344

Table 4. Detection Costs with Various Dynamic Threshold Functions (Initial Threshold = 0.05)

Additionally, we evaluate the effect of the window size. Table 5 shows the results using various window sizes in Formula (3). The best detection cost, i.e., 0.012647, is achieved under window size 2. It also shows the efficiency of dynamic threshold and window size.

Window size	1	2	3	4
$C_{\det}$	0.012657	**0.012647**	0.012809	0.012942

Table 5. Detection Costs with Various Window Sizes Using Formula (3) (Initial Threshold = 0.05)

4 Event Clustering Using Event Words

The co-reference chains in the above approach considered those features, such as person name, organization name, location, temporal expression and number expression. However, the important words "*black box*" or "*rescue*" in an air crash event are never shown in any co-reference chain. This section introduces the concepts of event words. Topic and event words were applied to topic tracking successfully (Fukumoto and Suzuki, 2000). The basic hypothesis is that an event word associated with a news article appears across paragraphs, but a topic word does not. In contrast, a topic word frequently appears across all news documents. Because the goal of event clustering is to extract all the events associated with a topic, those documents belonging to the same topic, e.g., China Airlines Air Accident, always have the similar topic words like “China Airlines”, “flight 611”, “air accident”, “Pen-Hu”, “Taiwan strait”, “rescue boats”, etc. Topic words seem to have no help in event clustering. Comparatively, each news article has different event words, e.g., “emergency command center”, “set up”, “17:10PM”, “CKS airport”, “Commander Lin”, “stock market”, “body recovery”, and so on. Extracting such keywords is useful to understand the events, and distinguish one document from another.

The postulation by Fukumoto and Suzuki (2002) is that 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. The former tells if a word appears across paragraphs (documents), and the latter tells if a word appears frequently. Event words are extracted by using these two values. Formula (5) defines a weight of term t in the i-th story.

$$Ws_{it} = \frac{TFs_{it}}{Max_j(TFs_{ij})} \times \log \frac{N}{Ns_t} \tag{5}$$

where TFs_{it} denotes term frequency of term t in the i-th story, N is total number of stories, and Ns_t is the number of stories where term t occurs.

Besides term weight in story level, Wp_{it} defines a weight of term t in the i-th paragraph. Formulas (6) and (7) define dispersion value and deviation value, respectively.

$$DispS_t = \sqrt{\frac{\sum_{i=1}^{m}(Ws_{it} - mean_t)^2}{m}} \tag{6}$$

$$DevS_{it} = \frac{(Ws_{it} - mean_t)}{DispS_t} \times \alpha + \beta \tag{7}$$

Where, $mean_t$ is average weight of term t in story level. Similarly, $DispP_t$ and $DevP_{jt}$ are defined in the paragraph level. The dispersion value of term t in the story level denotes how frequently term t appears across m stories. The deviation value of term t in the i-th story denotes how frequently it appears in a particular story. Coefficients α and β are used to adjust the number of event words. In our experiments, 20 event words are extracted for each document. In such a case, (α, β) is set to (10, 50) in story level and set to (10, 25) in paragraph level, respectively.

Formula (8) shows that term t frequently appears across paragraphs rather than stories. Formula (9) shows that term t frequently appears in the i-th story rather than paragraph P_j. An event word is extracted if it satisfies both formulas (8) and (9).

$$DispP_t < DispS_t \tag{8}$$

$$DevP_{jt} < DevS_{it} \quad \text{for all } P_j \text{ such that } P_j \in S_i \tag{9}$$

Below shows the event clustering using event words only. At first, we extract the event words of each news article using the whole news collection. For each sentence, we then compute the number of event words in it. After sorting all sentences, the designated number of sentences are extracted according to their number of event words. In the experiments, we use different window sizes to study the change of detection cost after introducing event words. Table 6 shows the experimental results under the same threshold (0.005) and test collection mentioned in Section 3.

Window size	1	2	3	4
C_{det}	0.011918	0.011842	**0.011747**	0.011923

Table 6. Detection Costs with Event Words and Various Window Sizes

The results in Table 6 are much better than those in Table 5, because inclusion of event words selects more informative or representative sentences or paragraphs. The more informative feature words documents have, the more effectively documents of one event can be distinguished from those of another. In other words, the similarities of documents among different events become smaller, so that the documents cannot be assigned to the same cluster easily under the higher threshold, and the best performance is shifted from window size 2 to window size 3.

5 Event Clustering Using Both Co-reference Chains and Event Words

According to the above experimental results, it is evident that either co-reference chains or event words are useful for event clustering on streaming news. As co-reference chains and event words are complementary in some sense, we further examine the effect on event clustering using both of them. Thus, two models called summation model and two-level model, respectively, are proposed. The summation model is used to observe the summation effect using both the co-reference chains and the event words on event clustering. On the other hand, the two-level model is used to observe the interaction between co-reference chains and event words.

5.1 Summation Model

In summation model, we simply add the scores for both co-reference chains and event words, which are described above respectively to be the score for each sentence in the news document. At first, we extract the event words of each news article using the whole news collection described in Section 3. For each sentence, we then compute the number of event words in it, and add this count to the number of co-reference chains it covers. The iterative procedure specified in Section 2 extracts the designated number of sentences according to the number of event words and co-reference chains.

Table 7 summarizes the experimental results under the same test collection mentioned in Section 3. The experiments of summation model show that the best detection cost is 0.011603. Comparing the best result with those in Tables 5 and 6, the detection costs are decreased 9% and 2%, respectively.

Window size	1	2	3	4
Cdet	0.112233	**0.011603**	0.013109	0.013109

Table 7. Detection Costs Using Summation Model

5.2 Two-level model

By comparing the experimental results described in Section 3 and 4, we noticed that the event word factor seems more important than the co-reference factor on event clustering of news document. Moreover, from the summation model we only know that both factors are useful on event clustering. In order to make clear which factor is more important during event clustering of news documents, 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. For example, we use the score function and the sentence selection algorithm described in Section 3 first, when there is a tie during sentence selection. Then we use the score function described in Section 4 to decide which sentence is selected from those candidate sentences, and vice versa. Thus, two alternatives are considered. Type 1 model uses the event words sentence selection algorithm described in Section 4 to select the representative sentences from each document, the co-reference chains are used to solve the tie issue. In contrast, type 2 model uses the co-reference chains sentence selection algorithm described in Section 3 to select the representative sentences for each documents and use event words to solve the tie issue. Table 8 shows the experimental result under the same test collection as described in previous sections.

Window size	2	3	4	5
Type 1	0.012116	0.011987	**0.011662**	0.012266
Type 2	0.012789	**0.012674**	0.012854	0.012941

Table 8. Detection Costs Using Two level Models

The performance of type 1 outperforms that of type 2. This result conforms to those shown in Table 5 and Table 6. We can say that the effect of event words is better than the co-reference chains in event clustering. Furthermore, the best score of type 1 is also better than the best score of Table 6. Thus, the introduction of co-reference chains can really improve the performance of event clustering using event words. On the other hand, the introduction of event words in type 2 does not have such an effect. Moreover, to further examine the use of co-reference chain information and the event words in event clustering, a more elaborate combination, e.g., using mutual information or entropy, of the two approaches is needed.

6 Concluding Remarks

This paper presented an approach for event clustering on streaming news based on both co-reference chains and event words. The experimental results using event words only outperform the results using the co-reference chains only. Nevertheless, as to the combination of co-reference chains and event words in event clustering, the experimental results show that the introduction of co-reference chains can improve the performance of event clustering using event words much. To model the temporal behavior of event clustering of streaming news, a dynamic threshold setting using time decay function and spanning window size is proposed. The experimental results, using TDT's evaluation metric – say, detection cost, show that the dynamic threshold is useful. .

We believe that the improvement of multi-document co-reference resolution will have great impact on temporal event clustering. In order to further improve our performance in even clustering on streaming news, there are still future works needed to be studied:

(1) In order to verify the significance of the experimental results, statistical test is needed.

(2) Instead of hand-tagging method, we will introduce automatic co-reference resolution tools to create large scale test corpus and conduct large scale experiments.

(3) When the length of document is variable, the fixed number of representative sentences may lose many important sentences to degrade the performance of event clustering. The dynamic

number of representative sentences for each document according to its length is introduced.

(4) As the news stories are reported incrementally instead of being given totally in the on-line event clustering, the computation of event words is an important issue.

(5) Apply the extracted sentences for each document to generate event-based short summary.

References

Allan, James; Carbonell, Jaime; and Yamron, Jonathan (Eds) (2002) Topic Detection and Tracking: Event-Based Information Organization, Kluwer.

Azzam, S.; Humphreys, K; and Gaizauskas, R. (1999) "Using Coreference Chains for Text Summarization," Proceedings of the ACL Workshop on Coreference and Its Applications, Maryland.

Bagga, A. and Baldwin, B. (1998) "Entity-Based Cross-Document Coreferencing Using the Vector Space Model," Proceedings of the 36th Annual Meeting of ACL and the 17th International Conference on Computational Linguistics.

Cardie, Claire and Wagstaff, Kiri (1999) "Noun Phrase Co-reference as Clustering," Proceeding of the Joint Coreference on EMNLP and VLC

Chen, Hsin-Hsi and Ku, Lun-Wei (2002) "An NLP & IR Approach to Topic Detection," Topic Detection and Tracking: Event-Based Information Organization, James Allan, Jaime Carbonell, and Jonathan Yamron (Editors), Kluwer, pp. 243-264.

Chen, Hsin-Hsi; Kuo, June-Jei and Su, Tsei-Chun (2003) "Clustering and Visualization in a Multi-Lingual Multi-Document Summarization System," Proceedings of 25th European Conference on Information Retrieval Research, Lecture Notes in Computer Science, LNCS 2633, pp. 266-280.

Fiscus, Jonathan G. and Doddington, George R. (2002) "Topic Detection and Tracking Evaluation Overview," Topic Detection and Tracking: Event-Based Information Organization, James Allan, Jaime Carbonell, and Jonathan Yamron (Eds), Kluwer, pp. 17-32.

Fukumoto, F. and Suzuki, Y. (2000) "Event Tracking based on Domain Dependency," Proceedings of the 23rd ACM SIGIR 2000 Conference, pp. 57-64

Kuo, June-Jei; Wong, Hung-Chia; Lin, Chuan-Jie and Chen, Hsin-Hsi (2002) "Multi-Document Summarization Using Informative Words and Its Evaluation with a QA System," Proceedings of The Third International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, LNCS 2276, pp. 391-401.

Manmatha, R.; Feng, A. and Allan, James (2002) "A Critical Examination of TDT's Cost Function," Proceedings of the 25th ACM SIGIR Conference, pp. 403-404.

MUC (1998) Proceedings of 7th Message Understanding Conference, Fairfax, VA, 29 April - 1 May, 1998, http://www.itl.nist.gov/iaui/894.02/ related_projects/muc/ index.html.

Silber, H. Gregory and McCoy, Kathleen F. (2002) "Eficiently Computed Lexical Chains As an Intermediate Representation for Automatic Text Summarization." Journal of Association for Computational Linguistics, Vol.28, No.4, pp. 487-496.

Wong, Hong-Jia; Kuo, June-Jei and Chen, Hsin-Hsi (2001) "Headline Generation for Summaries from Multiple Online Sources." Proceedings of 6th Natural Language Processing Pacific Rim Symposium, November 27-29 2001, Tokyo, Japan, pp. 653-660.