Topic and Subject Detection in News Streams for Multi-document Summarization

Fumiyo Fukumoto[1], Yoshimi Suzuki[1] and Atsuhiro Takasu[2]
[1]*Interdisciplinary Graduate School of Medicine and Engineering, Univ. of Yamanashi, Kofu, Japan*
[2]*National Institute of Informatics, Chiyoda-ku, Japan*

Keywords: Topic, Subject, Multi-document Summarization, Moving Average Convergence Divergence.

Abstract: This paper focuses on continuous news streams and presents a method for detecting salient, *key* sentences from stories that discuss the same topic. Our hypothesis about key sentences in multiple stories is that they include words related to the target *topic*, and the *subject* of a story. In addition to the TF-IDF term weighting method, we used the result of assigning domain-specific senses to each word in the story to identify a subject. A topic, on the other hand, is identified by using a model of "topic dynamics". We defined a burst as a time interval of maximal length over which the rate of change is positive acceleration. We adapted stock market trend analysis technique, *i.e.*, Moving Average Convergence Divergence (MACD). It shows the relationship between two moving averages of prices, and is popular indicator of trends in dynamic marketplaces. We utilized it to measure topic dynamics. The method was tested on the TDT corpora, and the results showed the effectiveness of the method.

1 INTRODUCTION

With the exponential growth of information on the Internet, it is becoming increasingly difficult for a user to read and understand all the materials that is potentially of interest. Multi-document summarization is an issue to attack the problem. It differs from single document summarization in that it is important to identify differences and similarities across documents. This can be interpreted as a question of how to identify a topic and a subject in series of stories. Here, a topic is the same as TDT project: something that occurs at a specific place and time associated with some specific actions (Allan, 2003). A subject, on the other hand, refers to a *theme* of the story itself, *i.e.*, something a writer wished to express. Much of the work on summarization has applied statistical techniques based on word distribution to the target document (Lin and Hovy, 2002). Other approaches explore to use machine learning or graph-based ranking method (Marcu and Echihabi, 2002; Wan and Yang, 2008). Wan *et al.* proposed two models, the Cluster-based conditional Markov Random Walk model and the Cluster-based HITS model, both use the theme clusters in the document set (Wan and Yang, 2008). However, most of these approaches does not deal with the identification of a topic and a subject in series of stories.

This paper focuses on extractive summarization and presents a method for detecting key sentences from continuous news streams. We assume that a key sentence in multiple documents includes words related to the target topic, and the subject of each story. Moreover, we assume that the sense of the subject word is related to the domain where domain is the traditional text categorization sense. For example, the word "court" in the target topic "Pinochet trial" is related to the subject, "Pinochet appealed his arrest and a London court agreed," and it has a sense of judicature in the legal/criminal domain. Here, "court" has at least two senses in the WordNet; judicature and tennis court. If we can find that the "count" from a story "Pinochet trial" has a domain-specific sense, *i.e.*, judicature sense, we can identify the "court" to a subject word. In addition to the traditional term weighting method TF-IDF, we used the result of assigning domain-specific senses (ADSS) to identify subject words. On the other hand, a topic is identified by using a model of "topic dynamics". We defined a burst as a time interval of maximal length over which the rate of change is positive acceleration. We used Moving Average Convergence Divergence (MACD) to identify topic. MACD is a technique to analyze stock market trend. It shows the relationship between

Fukumoto F., Suzuki Y. and Takasu A..
Topic and Subject Detection in News Streams for Multi-document Summarization.
DOI: 10.5220/0004109901660171
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 166-171
ISBN: 978-989-8565-30-3

two moving averages of prices modeling bursts as intervals of topic dynamics, *i.e.*, positive acceleration.

2 SYSTEM DESIGN

2.1 Assignment of Domain-specific Senses for Subject Detection

We used the result of ADSS to identify subject words. We used the TDT corpus and WordNet 3.0 thesaurus. The TDT documents are classified into eleven topics (domains), such as "natural disasters" and "elections". We used these topics except for the topic MISC. To assign topics/domains of the TDT corpus to each sense of the word in WordNet, we first selected each sense of a word to the corresponding domain by using a text classification technique. For each sense s of a word w, we replace w in the training stories assigning to the topic t with its gloss text in WordNet. (hereafter, referred to as word replacement). If the classification accuracy of the topic t is equal or higher than that without word replacement, the sense s is regarded to be a domain-specific sense. However, the sense selection is not enough for ADSS. Because the number of words consisting gloss in WordNet is not so large. As a result, the classification accuracy with word replacement was equal to that without word replacement[1]. Then we scored senses by computing the rank scores.

1. Candidate Extraction

The first step to find domain-specific senses is to extract candidates. We divided TDT documents into two: training data to learn SVM model and test data to classify documents. For each topic, we collected a set of words W with high TF-IDF value from the TDT corpus. Let TS be a topic set, and S be a set of senses that the word $w \in W$ has. The candidates are obtained as follows:

1. For each sense $s \in S$, and for each $t \in TS$, we replace w in the training documents assigning to the topic t with its gloss text in the WordNet.
2. All the documents of training and test data are tagged by a part-of-speech tagger, stop words are removed, and represented as term vectors with frequency.
3. The SVM is applied to the two types of the training documents, *i.e.*, with and without word replacement, and classifiers for each topic are generated.
4. SVM classifiers are applied to the test data. If the classification accuracy of the topic t is higher than that without word replacement, the sense s of the word w is judged to be a candidate sense in the topic t.

The procedure is applied to all $w \in W$.

2. Scoring Senses by Link Analysis

The next procedure for ADSS is to score each candidate for each topic. We used the MRW model. Given a set of candidates C_t in the topic t, $G_t = (S, E)$ is a graph reflecting the relationships between senses in the candidate set. S is a set of vertices, and each vertex s_i in S is a gloss text assigned from the WordNet. E is a set of edges, which is a subset of $S \times S$. Each edge e_{ij} in E is associated with an affinity weight $f(i \to j)$ between senses s_i and s_j $(i \neq j)$. The weight is computed using the standard cosine measure between the two senses. Two vertices are connected if their affinity weight is larger than 0 and we let $f(i \to i) = 0$ to avoid self transition. The transition probability from s_i to s_j is then defined as follows:

$$p(i \to j) = \begin{cases} \frac{f(i \to j)}{|S|}, & \text{if } \Sigma f \neq 0 \\ \sum_{k=1} f(i \to k) & \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

We used the row-normalized matrix $U_{ij} = (U_{ij})_{|S| \times |S|}$ to describe G with each entry corresponding to the transition probability, where $U_{ij} = p(i \to j)$. To make U a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to $\frac{1}{|S|}$. The matrix form of the saliency score $Score(s_i)$ can be formulated in a recursive form as in the MRW model.

$$\vec{\lambda} = \mu U^T \vec{\lambda} + \frac{(1-\mu)}{|S|} \vec{e}. \quad (2)$$

where $\vec{\lambda} = [Score(s_i)]_{|S| \times 1}$ is the vector of saliency scores for the senses. $\vec{e}$ is a column vector with all elements equal to 1. μ is the damping factor, which we set to 0.85. The final transition matrix is given by the Eq. (3), and each score of the sense in a specific domain is obtained by the principal eigenvector of the matrix M.

$$M = \mu U^T + \frac{(1-\mu)}{|S|} \vec{e}\vec{e}^T. \quad (3)$$

The procedure is applied to all of the topics. We selected a certain number of words (senses) according to rank score as a subject word in a document.

[1] In the experiments, the classification accuracy of more than 50% of words has not changed.

2.2 Topic Detection

Topic Bursts

He *et al.* proposed a method to find bursts by using Moving Average Convergence/Divergence (MACD) histogram which was used in technical stock market analysis to detect bursts (He and Parker, 2010). MACD histogram refers to a difference between the MACD and its moving average. MACD is defined by Eq. (4).

$$hist(n_1,n_2,n_3) = MACD(n_1,n_2) - EMA(n_3)[MACD(n_1,n_2)] \quad (4)$$

$EMA(n_3)$ refers to n_3-day Exponential Moving Average (EMA). For a variable $x = x(t)$ which has a corresponding discrete time series $\mathsf{x} = \{x_t \mid t = 0,1,\cdots\}$, the n-day EMA is defined by Eq. (5).

$$EMA(n)[x]_t = \alpha x_t + (1-\alpha)EMA(n-1)[x]_{t-1} = \sum_{k=0}^{n} \alpha(1-\alpha)^k x_{t-k} \quad (5)$$

α refers to a smoothing factor and it is often taken to be $\frac{2}{(n+1)}$. $MACD(n_1,n_2)$ in Eq. (4) indicates the difference of n_1-day and n_2-day exponential moving averages, *i.e.*, $EMA(n_1) - EMA(n_2)$. We applied the model to detect topic words.

Topic Detection

The procedure for topic detection is illustrated in Figure 1. Let A be a set of documents to be summarized. A set of topic words TS are detected as follows:

1. Create MACD histogram where X-axis refers to a period of time of length T, and Y-axis denotes the frequency of documents concerning to the target topic. Hereafter, referred to as correct histogram, as shown in Figure 1.
2. Each term in the TDT corpus is weighted by using TF-IDF scheme. For each target topic, terms within the documents assigning to the target topic are sorted in the descending order to make a term list.
3. Given the number of k, we extracted the topmost k terms from the term list. For each term, we applied the following procedures.
   (a) Create MACD histogram where X-axis refers to a period of time of length T, and Y-axis denotes bursts. Hereafter, referred to as bursts histogram, as shown in Figure 1.

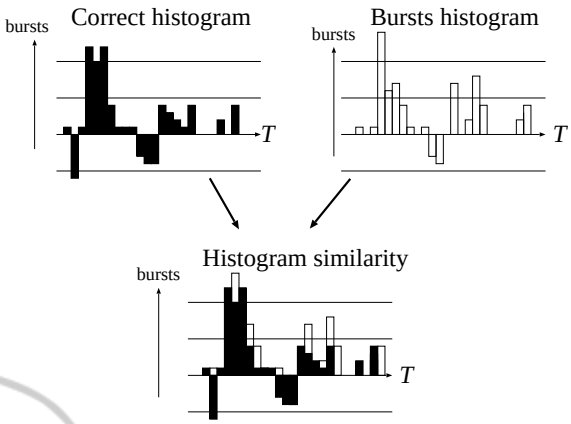

Figure 1: Similarity between correct and bursts histograms.

   (b) As illustrated in the bottom of Figure 1, compute similarity between correct and bursts histograms by using Bhattacharyya distance, $\rho(p,q) = \sum_{i=1}^{T} \sqrt{p_i q_i}$ where p and q are a normalized distance of correct histogram and burst histogram, respectively[2]. p_i refers to the frequency of documents that arrive in time i, and q_i indicates bursts of topic in time i. If the value of $\rho(p,q)$ is larger than a certain threshold value, the term t is regarded as a topic word.

In the procedure (b), we assume that burst histogram of the term t is close to the correct histogram if t is a topic term. Because burst histogram obtained by procedure (a) refers to a burst of a topic t concerning to a set of documents A. Thus, we can assume that it is similar to the histogram obtained by using a frequency of A concerning to the target topic.

2.3 Sentence Extraction

Each sentence concerning to the target topic is represented using a vector of frequency weighted words that can be subject or topic words. Similar to the procedure of ADSS, we used the MRW model to compute the rank scores for the sentences, *i.e.*, given a document set D, a graph G consists of a set of vertices S and each vertex s_i in S is a sentence in the document set. After the saliency scores of sentences have been obtained, choose a certain number of sentences according to the rank score into the summary.

[2]We tested Bhattacharyya distance, histogram intersection and KL-distance to obtain similarities. We reported only the result obtained by Bhattacharyya distance as it was the best results among them.

Table 1: Topics assigned to categories in the TDT.

Category	Topics
Elections	U.S. Mid-Term Elections
Scandals/hearings	Olympic Bribery Scandal
Legal/criminal	Pinochet Trial
Natural disasters	Hurricane Mitch
Accidents	Nigerian Gas Line fire
Violence or war	Car Bomb in Jerusalem
Science and discovery	Leonid Meteor Shower
Finances	IMF Bailout of Brazil
New laws	Anti-Doping Proposals
Sports	Australian Yacht Race

Table 2: The number of candidate senses.

Category	Doc	Total	Cand.(%)
Elections	518	3,837	1,970(51.3)
Scandals/hearings	108	3,566	1,775(49.7)
Legal/criminal	860	3,396	1,805(53.1)
Natural disasters	286	3,496	1,851(52.9)
Accidents	82	2,828	916(32.3)
Violence or war	234	3,495	1,741(49.8)
Science and discovery	186	3,955	1,969(49.7)
Finances	540	3,428	1,945(56.7)
New laws	7	3,304	1,646(49.8)
Sports news	326	3,428	1,957(57.0)

3 EXPERIMENTS

We used the TDT3 corpus which comprises a set of eight English news sources collected from October to December 1998. It consists of 34,600 stories. A set of 60 topics are defined for evaluation in 1999, and another 60 topics for evaluation in 2000. Of these topics, we used 78 topics, each of which is classified into 10 categories. Table 1 illustrates categories and some examples of topics assigned to these categories.

3.1 Assignment of Domain-specific Senses

We divided TDT documents into two: training and test data in text classification. The size of training data for each category is two-third of documents, and the remaining is test data. All documents were tagged by Tree Tagger (Schmid, 1995). For each category, we collected the topmost 500 noun words with high TF-IDF weight from the TDT3 corpus. We used WordNet 3.0 to assign senses. Table 2 shows the number of training documents, the total number of senses, and the number of candidates senses (Cand.) that the classification accuracy of each category was higher than the result without word replacement. We used these senses as an input of the MRW model.

Table 3: The result against SFC resource.

Cat	ADSS	SFC	SFC & TDT	Recall
Finances	390	125	81	0.648
New law	358	1,628	193	0.437
Science	389	671	176	0.699
Sports	395	1,947	8	1.000

There are no existing sense-tagged data for these 10 categories that could be used for evaluation. Therefore, we selected a limited number of words and evaluated these words qualitatively. To this end, we used the Subject Field Codes (SFC) resource (Magnini and Cavaglia, 2000) annotating WordNet 2.0 synsets with domain labels. The SFC consists of 115,424 words assigning 168 domain labels with hierarchy. It contains ”finances”, ”laws”, ”science” and ”sports” labels. We used these four labels and its children labels in a hierarchy, and compared the results with SFC resource. The results are shown in Table 3. ”ADSS” shows the number of senses assigned by our approach. ”SFC” refers to the number of senses appearing in the SFC resource. ”SFC & TDT” denotes the number of words (senses) appearing in both SFC and the TDT corpus. We note that the corpus we used was TDT corpus, while SFC assigns domain labels to the words appearing in the WordNet. Therefore, we used recall as the evaluation measure where it refers to the number of senses matched in our approach and SFC divided by the total number of senses appearing in both SFC and TDT. ”Recall” in Table 3 refers to the best performance among the varying number of senses according to the rank scores. As we can see from Table 3 that word replacement improved text classification performance as the former was 0.06 F-score, while that of the latter was only 0.01. One reason is the length of the gloss text in the WordNet; the average length of gloss text assigned to ”law” was 5.75, while that for ”sports” was 8.96. The method of assigning senses depends on the size of gloss text in the WordNet. Efficacy can be improved if we can assign example sentences to WordNet based on corpus statistics. This is a rich space for further exploration.

It is interesting to note that some senses of words that were obtained correctly by our approach did not appear in the SFC resource because of the difference in WordNet version, *i.e.*, we used WordNet 3.0 and the TDT corpus for ADSS, while SFC is based on WordNet 2.0. Table 4 illustrates some examples obtained by our approach but that did not appear in the SFC. These observations support the usefulness of our automated method.

For evaluating subject detection, we randomly selected one topic for each category, and manually checked whether the words assigning domain-specific senses are the subject or not. Table 5 shows the re-

Table 4: Some examples obtained by ”ADSS”.

Cat	Example of words and their senses	
New law	fire:	intense adverse criticism
	break:	an escape from jail
Sports	era:	(baseball) a measure of a pitcher's effectiveness

Table 5: Performance of subject detection.

Cat	ADSS	Subject	F	F(TF-IDF)
Finances	86	88	0.678	0.302
New law	63	79	0.577	0.208
Science	72	102	0.609	0.319
Sports	92	132	0.705	0.412

sult of subject detection. ”ADSS” refers to the number of subject words identified by our approach and also appeared in the documents assigning to the target topic. ”Subject” denotes the number of correct subject words identified by two humans[3], and ”F” shows F-score. We compared the result obtained by our approach with simple term weighting method, TF-IDF. ”F(TF-IDF)” shows F-score obtained by using TF-IDF weighting method. As can be seen clearly from Table 5, the results obtained by our approach were much better than those of a simple term weighting, TF-IDF method.

3.2 Sentence Extraction

Finally, we report the results of sentence extraction. Because of manual creation of the evaluation data, we used 32 out of 78 topics which have less than 646 sentences in documents in the experiment. The evaluation is made by two humans. We set the extraction ratio to 30%, and compared our method with the following three approaches to examine how the results of subject and topic detection affect sentence extraction. The average number of sentences with the extraction ratio of 30% was 42.5 sentences.

1. Apply MRW model to noun words (Noun)
   The method applied the MRW model to the sentences consisting of noun words.
2. Apply MRW model to topic words (Topic)
   The method applied the MRW model to the results of MACD method.
3. Apply MRW model to subject words (Subject)
   In contrast to “Apply MRW model to subject words”, the method applied the MRW model to the results of ADSS.

For each of the four methods including our approach, we divided 32 topics into two: 10 topics to train the optimal number of subject and topic words, and the remaining 22 topics to test sentence extraction performance by using the estimated number of subject and topic words. The best performance obtained by ADSS (Subj) using training data was 0.332 ROUGE-1 score and the number of subject words was 10 per topic. Similarly, the best performance by our approach (Subj & Topic) was 0.442 ROUGE-1 score and the number of subject and topic words were 10 and 5, respectively. Therefore, we used the topmost 10 subject and 5 topic words, and evaluated each method by using the test data. The results are shown in Table 6. In Table 6, ”Max” and ”Min” denote the maximum and minimum ROUGE-1 score, respectively. ”Ave” refers to the average score obtained by using 22 topics.

Table 6: ROUGE-1 score for 22 topics.

Method	Max	Min	Ave
Noun	0.428	0.180	0.258
Subject	0.485	0.202	0.269
Topic	0.529	0.362	0.421
Subject & Topic	0.750	0.318	0.485

As shown in Table 6, the use of subject words only did not contribute sentence extraction performance, as the average ROUGE-1 score was 0.269 and it was not significant improvement over the baseline, *i.e.*, the method applied MRW model to the sentences consisting noun words. The use of topic words only contributes sentence extraction performance compared to the use of subject words only. Moreover, the results obtained by using both subject and topic words, *i.e.*, our approach attained at 0.485 averaged ROUGE-1 score, and we found that it always outperforms, regardless of how many number of documents (sentences) were used. These results clearly support the usefulness of our method.

Figure 2 illustrates the topmost three sentences extracted by each method. The topic is ”Ukraine Mining Accidents”. ”◯” indicates that the system and human judges agree. Words marked with ”{}” and the underlined words refer to a subject word and a topic word identified by the system, respectively. Figure 2 shows that terms such as ”Ukraine”, ”mines”, and ”accident” consisting topic name are correctly extracted by ”Topic” and ” Subject & Topic” methods. Similarly, ”Donetsk” and ”region” appearing a particular document is correctly extracted as a subject word by ”Subject” method. Our method correctly extracted salient sentences, while the methods used subject or topic only were not perfect extraction. Moreover, the results obtained by ”Subject” shows that sentences 1 and 2 are similar contents, *i.e.* the extracted sentences include redundancy information, while those obtained by ”Subject & Topic” shows that sentences 1 and 2 are subsumption relationship, *i.e.*, sentence 2 in-

[3]The classification is determined to be correct if two human judges agrees.

Method: Noun
1 Pieces of coal fell in a deep mine shaft in eastern Ukraine on Monday, killing a miner, officials said.
2 Ukraine's cash-strapped mines lack the funds needed to improve safety conditions and modernize equipment, which has led to a steady increase in work-related fatalities in recent years.
③Two mine workers were killed in separate accidents in eastern Ukraine, pushing this year's number of mine fatalities to more than 300, officials said Wednesday.

Method: Subject
1 Two {miners} were killed in {Ukraine's} crumbling {coal} {mines} over the weekend, bringing to at least 305 the number of {coal} industry {workers} who have died on the job so far this {year}, officials said Monday.
②Two {mine} {workers} were killed in separate accidents in eastern {Ukraine}, pushing this {year's} number of {mine} fatalities to more than 300, officials said Wednesday.
③The worst single accident this {year} was in April, when a methane gas blast killed 63 {workers} at a {mine} in the eastern {Donetsk} {region}.

Method: Topic
1 The accident occurred at the depth of 540 meters (1,782 feet) at the Olkhovatska mine in the coal-rich Donetsk region, the Emergency Situations Ministry reported.
②The death toll in coal mine accidents in Ukraine has been rising for years as the cash-strapped government has been unable to modernize deteriorating mines and improve safety conditions.
③Work and safety conditions at the former Soviet republic's coal mines have deteriorated in recent years as the cash-strapped government has failed to provide sufficient funds to support the largely unreformed industry.

Method: Subject & Topic
①Two {mine} {workers} were killed in separate accidents in eastern {Ukraine}, pushing this {year's} number of {mine} fatalities to more than 300, officials said Wednesday.
②A total of 301 people have died in {mine} accidents in {Ukraine} so far in 1998, about 100 more than over the same period last {year}, according to government statistics.
③The number of {miners} killed on the job has been increasing steadily in recent {years} because {mines} lack the funds needed to modernize equipment and improve safety conditions.

Figure 2: Topmost three sentences extracted by each method ("Ukraine Mining Accidents").

cludes additional information, "about 100 more than over the same period last year". These observations again clearly support the usefulness of our method.

4 CONCLUSIONS

We have developed an approach to multi-document summarization from continuous news streams. The results showed the effectiveness of the method. Future work will include: (i) comparison to other topic models such as hierarchical Pachinko Allocation Model (Mimno et al., 2007) and Two-Tiered Topic Model (Celikylmaz and Hakkani-Tur, 2011), (ii) comparison to other term weighting methods such as Information Gain and χ^2 statistics, and (ii) applying the method to other data such as DUC2004 and DUC2007 for quantitative evaluation.

REFERENCES

Allan, J., editor (2003). *Topic Detection and Tracking*. Kluwer Academic Publishers.

Celikylmaz, A. and Hakkani-Tur, D. (2011). Discovery of Topically Coherent Sentences for Extractive Summarization. In *Proc. of the 49th ACL*, pages 491–499.

He, D. and Parker, D. S. (2010). Topic Dynamics: An Alternative Model of Bursts in Streams of Topics. In *Proc. of the 16th ACM SIGKDD*, pages 443–452.

Lin, C.-Y. and Hovy, E. H. (2002). From Single to Multi-Document Summarization: A Prototype System and its Evaluation. In *Proc. of the 40th ACL*, pages 457–464.

Magnini, B. and Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. In *In Proc. of the 2nd LREC*.

Marcu, D. and Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In *In Proc. of the 40th ACL*, pages 368–375.

Mimno, D., Li, W., and McCallum, A. (2007). Mixtures of Hierarchical Topics with Pachinko Allocation. In *In Proc. of the 24th ICML*, pages 633–640.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proc. of the EACL*.

Wan, X. and Yang, J. (2008). Multi-Document Summarization using Cluster-based Link Analysis. In *Proc. of the 31st ACM SIGIR*, pages 299–306.