CS474 Text Mining Project

Issue Trend Analysis & Issue Tracking

Junmo Cho
Electrical Engineering
KAIST
Republic of Korea
junmokane@kaist.ac.kr

Byoungjun Kim
Electrical Engineering
KAIST
Republic of Korea
braian98@kaist.ac.kr

ABSTRACT

This is the report of CS474 Text Mining Project. There are two tasks that we have to handle. The first one is **Issue Trend Analysis**. This task is to find the top 10 trending issues from documents. The second one is **Issue Tracking**. First, we need to pick two issues from the top 10 trending list from the first task. Each issue contributes to each sub-task (Issue Tracking consists of two sub-tasks). One is **On-Issue Event Tracking**. For this task, we need to describe the change of events related to the issue with a timeline with some information. Another one is **Related-issue Event Tracking**. For this task, we need to extract and describe the related events with some information that are not directly linked to the issue chosen, but topically related.

Since this is open problem, we will define some terms related to this project to handle the problem logically. For the Issue Trend Analysis, we basically used LDA which is the famous topic modeling techniques. However, extracting top issues using only LDA is not easy, because LDA just gives topic top keywords, and each topic includes broad and various documents. To narrow down to certain issue, with topics given by LDA, we used another clustering method DBSCAN for documents in each topic. For some topic, we got clusters by DBSCAN and chose the cluster that has the largest documents in the topic. From that, we extracted the issues with HMM (Hidden Markov Model) language model. Repeating this process, we have issue for each topic, then we sort those issues with the corresponding number of documents. From sorted issues, we just picked top 10 for the top 10 trending issues.

From Issue Trend Analysis, we picked two issues. One issue is for On-Issue Event Tracking, and another issue is for Related-issue Event Tracking. For On-Issue Event Tracking, we first observed the documents from that issue in timeline. To extract the events with timeline from those documents, dividing documents that represent similar event is needed. We approached this dividing problem with NER (Named Entity Recognition). By looking through entity such as PERSON, ORGANIZATION from documents with timeline, if documents have overlapping entity values, we classified them as same event. Now, we have divided documents and its NER values. Final thing is to extract the event from each divided documents. This is simply done by using HMM language model. After extracting events and NER values, we applied some general rules to reduce noise which will be discussed later.

For Related-issue Event Tracking, from Issue tracking, we've got several clusters for one topic by using DBSCAN method. Since each cluster is included in one topic, this is exactly what we want. Each clusters will be topically related, but not directly related (This needs assumption that LDA performed well). So, for each clusters, we applied NER for information extraction, and HMM language model for event extraction.

To logically support our approach, we evaluated methods we used with evaluation methods. For LDA, we used perplexity, and coherence value to choose appropriate topic numbers. For DBSCAN, we evaluated the performance of the method with alternatives of Dunn Index. For HMM and NER, there is no ground truth data for evaluating those methods. So, we just judged the result of these methods by ourselves (human judgement).

KEYWORDS

text mining, LDA, DBSCAN, HMM, NER

1 Problem Statements

For logical approach, we need to define some terms.

1.1 Issue Trend Analysis

We defined issue as key phrase which represents cluster of documents. We used LDA topic modeling to extract topic keywords and by looking at results of topic keywords, we found out that topic keywords are broad term so it can't be used to specify specific issue. However, after divide cluster from topic, those clusters are more specific and we thought it could be used as issue. Size of the cluster represents importance of such issue, therefore we defined top trending issue as key phrase of cluster in order of size.

1.2 On-issue event tracking

As mentioned above, we extracted top 10 trending issue as mean of LDA and clustering. For on issue tracking, we sort biggest cluster by time correspond to each topic and cluster time continuing document if information of person is same. If key information of

document doesn't change through time, it means each document represents similar event. Therefore, we cluster time continuing document if information of person is same.

1.3 Related-issue event tracking

We defined related issue as key phrase each cluster in same topic except the largest cluster. Since topic represents broad term, which indicates each cluster in same topic are related. However, event of each clusters is not directly related. Therefore, for related issue tracking we extracted key phrase each cluster in same topic except the largest cluster, since largest cluster represents issue.

2 Method

In this part, we explain how methods are applied in detail. All programs are in python. We used many text mining libraries such as nltk, spacy, genism, and etc.

2.1 Issue Trend Analysis

Overall process of issue tracking in detail is as follows:
document classification (2015, 2016, 2017) – text preprocessing – LDA – DBSCAN – HMM – Selecting Top 10 Issues

2.1.1 document classification. We just classify the documents with corresponding years. We have three sets of documents (2015, 2016, 2017) in the beginning. Let's call them set A, B, C.

2.1.2 text preprocessing. Select one of documents set (e.g. A). For text preprocessing, following has been done in order.

parsing sentence into words – remove stop words – make n-gram – lemmatization with POS tagging – document term matrix

parsing sentence into words – given document set A, parse them into words. letter such as punctuations, ',', '.' are removed.

remove stop words – given set of words, remove stop words. We used list of stop words in nltk library.

make n-gram – given set of words, we made all cases of bigrams and trigrams. If generated bigrams or trigrams appear more than 5 times, we included them in the set of words.

lemmatization with POS tagging – givens set of words, we applied POS tagging for each words. nltk POS tagger is used. After POS tagging, lemmatization is applied for each words with tags. nltk WordNetLemmatizer is used.

document term matrix – given set of words, we formed document term matrix. However, we removed the words with high frequency and low frequency. If the word has document frequency higher than 0.8 or has term frequency lower than 3, we removed the word from the set of words. After this, we made document term matrix. sklearn CountVectorizer is used.

2.1.3 LDA. Given document term matrix, we applied LDA. However, to choose appropriate topic numbers, evaluation is needed. We calculated perplexity and coherence values for the set of number of topics (Here, we used topic number as 26 for all years). We chose the number of topics with the highest coherence value. After LDA, we have n topics $\{t_1, t_2, \dots, t_n\}$. This set is sorted in number of documents.

$\#\ documents(t_1) > \#\ documents(t_2) > \cdots > \#\ documents(t_n)$

2.1.4 DBSCAN. Given topic t_i, we applied DBSCAN on the set of documents that has the largest connection to topic t_i. We used title of documents for DBSCAN. For preprocessing, we mapped each title into vectors. Libraries tdqm, and "en_core_web_lg" from spacy were used for vectorization. After vectorization, DBSCAN is applied. However, to choose appropriate epsilon, evaluation is needed. We calculated alternative evaluation method of Dunn Index to choose epsilon. Alternative method is to choose epsilon that has the highest number of clusters. The reasons why we used alternative method and why it is okay to choose epsilon this way are explained in Deep Analysis part. After DBSCAN, for topic t_i, we have n_i document clusters $\{c_{i1}, c_{i2}, \dots, c_{in_i}\}$. From document clusters, we picked the cluster c_{ik_i} that has the largest number of documents. With this cluster, we will extract the issue.

2.1.5 HMM. Given set of clusters $\{c_{1k_1}, c_{2k_2}, \dots, c_{nk_n}\}$. This set consists of largest clusters from each topic. We extracted issues from each cluster by applying HMM on documents title. We used bigram model which is simple but result is reasonable. From title, we generated bigram model, and we generated random sentence. After generating random sentence, we calculated score of sentence in mean of conditional probability distribution of each word in sentence. We generated several random sentences and calculated score of each sentence and sorted in order. From these sentences, we picked top 10 generated sentences that explain the overall topic. Then, we picked central sentence from top 10 sentences by vectorization (same vectorization method discussed in DBSCAN preprocessing). The selected central sentence is the one issue. We do this process for each cluster in $\{c_{1k_1}, c_{2k_2}, \dots, c_{nk_n}\}$. Then we have n issues extracted $\{I_1, I_2, \dots, I_n\}$.

2.1.6 Selecting Top 10 Issues. We have n issues $\{I_1, I_2, \dots, I_n\}$. Also, we have corresponding clusters $\{c_{1k_1}, c_{2k_2}, \dots, c_{nk_n}\}$, and topics $\{t_1, t_2, \dots, t_n\}$. We introduce two ways of selecting top 10 issues. The first method is to simply pick $\{I_1, I_2, \dots, I_{10}\}$. This means we pick top issues based on number of documents in each topic (refer 2.1.3). The second method is to pick top issues based on number of documents in clusters $\{c_{1k_1}, c_{2k_2}, \dots, c_{nk_n}\}$. This will rearrange the orders in clusters as $\{c_{l_1}, c_{l_2}, \dots, c_{l_n}\}(\ \#\ documents(c_{l_1}) > \#\ documents(c_{l_2}) > \cdots > \#\ documents(c_{l_n}))$. Then we pick $\{I_{l_1}, I_{l_2}, \dots, I_{l_{10}}\}$. Let's denote first method as method A, second method as method B.

2.2 On-Issue Event Tracking

Overall process of on-issue event tracking in detail is as follows:
sort by time stamp – information extracting using NER – noise elimination - clustering document – event extraction by HMM
Let's denote largest cluster of topic t_i be c_{ik_i}. On section 2.1, we already clustered c_{ik_i} so in here, we will assume c_{ik_i} already exist.

2.2.1 sort by time stamp. We extracted time stamp information from each document of c_{ik_i}, then sorted by its order.

2.2.2 information extracting using NER. We used Standford NER through document body of c_{im} to extract information about person, organization and location.

2.2.3 noise elimination. After NER, there exist noise, such as person name or location in organization, different expression of person name and special character may be contained in information. Therefore, we eliminated those noises. If information in organization is already in person or location, we removed it. Also, we removed information containing special character. For person information, if information A is substring of information B, then we removed A from information (e.g. Park, Park Gyunhye -> remove Park).

2.2.4 cluster document. From information extracted by Standford NER, we scan through time continuing order and merge documents in c_{ik_i} if two documents (adjacent by time) have same person entity value. Reason for only considering person information is because person is principal agent of event. For example, let c_{ik_i} consists of documents set $\{d_1, d_2, \dots, d_m\}$ in timeline order. If person entity values of d_1 and that of d_2 have common value, merge them. So set becomes $\{d_{1,2}, d_3, \dots, d_m\}$. We repeat this process. Final form will be as follows.

$$\{d_{1,2,\dots,m_1}, d_{m_1+1,\dots,m_2}, \dots, d_{m_{k-1}+1,\dots,m_k}\}(m_k = m)$$

Reform above set as

$$\{ct_1, ct_2, \dots, ct_k\}$$

2.2.5 event extraction by HMM. We run HMM on data title of ct_j, and its result represent j^{th} time event of corresponding issue, which is HMM result of c_{im}. HMM works as mentioned in 2.1.5.

2.3 Related-issue Event Tracking

Overall process of related-issue event tracking in detail is as follows:

event extraction by HMM – information extraction using NER – noise elimination

Let's denote cluster of t_i be $(c_{i1}, c_{i2}, \dots, c_{in_i})$. If LDA performed well.

2.3.1 event extraction by HMM. We first ran HMM to each cluster c_{i1} to c_{in_i}. Process of HMM is same as mentioned in 2.1.5. Result of HMM represents related event.

2.3.2 information extraction using NER. We used NER as mentioned in 2.2.2 to extract information of c_{i1} to c_{in_i}. Result of NER represents information of each related event.

2.3.3 noise elimination. We eliminated noise of NER result as mentioned in 2.2.3.

3 Result

In this part, we show the result of our program.

3.1 Issue Trend Analysis

Ground truth – Man manually formed top 10 issue from scanning through title of document

X – GT does not contain single topic. Failure case

Result is in Appendix.

3.2 On-Issue Event Tracking

Below shows the result of on-issue event tracking. For information extraction, since there are a few events, we only posted one example, but whole result can be found at Appendix.

Example

Issue : S.Korea Japan to hold first security talks this week.

Event : Seoul, Tokyo to hold high-level economic talks, S.Korea, U.S. to hold cyber-security talks, U.S. defense chief to visit S.Korea : source, S.Korea, Japan to hold series of defense, foreign official's meetings, S.Korea, Japan to hold first security talks in 5 yrs., S.Korea, U.S. hold annual defense talks, S.Korea, Japan to hold first defense minister's talks, S.Korea, Japan to reopen finance minister's talks on end, S.Korea, Australia hold working-level defense talks, Korea, Japan in talks to resume working-level defense dialogue channel, S.Korea joint talks of foreign, defense ministers, S.Korea, U.S. to hold defense talks, Korea, Japan defense chief to meet in Seoul, S.Korea, China to hold talks on sea boundary

Detail information :

Event : U.S. defense chief to visit S.Korea: source

Person : Kim Min-Seok, Ashton Carter, John Kerry, Martin Dempsey

Organization : United States Joint Chiefs of Staff, Adm, Terminal High-Altitude Area Defense, THADD

Place : North Korea, Alaska, Washington, Japan, Northeast Asia, Afghanistan, Seoul, South Korea, California, China, The United States

Ground truth

Issue : S.Korea talk with other country

Event : S.Korea and Japan high-level economic talks, S.Korea, U.S. cyber security talks, U.S. defense chief visit, S.Korea, Japan security talks, S.Korea, U.S. defense talk, S.Korea, Japan defense minister's talks, S.Korea, Australia working-level defense talks, S.Korea, Japan working-level defense talks, S.Korea, Japan director-level defense talks, S.Korea, Australia joint talks of foreign, defense minister, S.Korea, U.S. defense talks, S.Korea, Japan defense chief meet, S.Korea, China talks on sea boundary

Detail information :

Event : U.S. defense chief visit

Person : Kim Min-Seok, Han Min-Koo, Ashton Carter, Park Geun-hye, Martin Dempsey, Choi Yun-hee,

Organization :

Place : North Korea, United States, South Korea, Seoul, Washington, Japan, China, Afghanistan, California, Alaska

3.3 Related-issue Event Tracking

Below shows the result of on-issue event tracking. For information extraction, since there are a few events, we only posted one example, but whole result can be found at Appendix.

Example

Issue : S.Korea's visit N.Korea to hold security dialogue

Event : N.Korea fishermen back home, Koreas to hold high-level talks, U.S. denies asking President Park to attend China ceremony,

N.Korea for inter-korean exchanges, Chinese tours to lift Ebola restrictions: tour firm
Detail information :
Event : N.Korea fishermen back home
Person : surnamed Lee, Kim Jun-wook, Kim Kuk-gi, Shin Hyon-hee, surnamed Jin, Joo Won-moon
Organization : Yonhap, Red Cross, Unification Ministry, Three North Korean, Korean Central New Agency, New York University, Unification, Choe, Coast Guard, KCNA
Place : Pyongyang, Panmunjeom, South Hamkyong Province, Ulleungdo, Seoul, China, East Sea, Wonsan, Panmunjom, United

Ground truth
Issue : S.Korea talk with Japan, U.S., China about security about N.Korea nuclear
Event : N.Korea fishermen back home, Two Koreas high level talks, Park attending China ceremony, Inter-Korean exchanges, N.Korea and China Ebola
Detail information :
Event : N.Korea fishermen back home
Person : Kim Kuk-gi, Choe Chun-gil, Joo Won-moon, Shin Hyon-hee
Organization : Unification Ministry, New York University, Central News Agency, KCNA, Korean Central News Agency
Place : North Korea, South Korea, Panmunjom, Inter-Korean border, Pyongyang, China, Seoul, South Hamkyong Province, Ulleungdo, Wonsan, East Sea, United States

4 Analysis

In this part, we will describe analysis of result and reasoning of our method. Also, for each part, we will introduce evaluation tactics for issue trend analysis and NER result.

4.1 Result Analysis

By comparing method A and method B, method B is more reasonable since top 10 issue is more diverse and those 10 issue are more familiar. For issue extracting, since we used bigram HMM model which can possibly generate sentences which are not related to document or grammatically wrong. However, in most of the case HMM generated sentence was closely related to ground truth. Also, we restricted the length to be generated to be bigger than 2/3 of average title length, therefore, generated sentence can be too specific. For example, in figure 1's 3th issue, generated sentence is 'S.Korea, Japan to hold first security talks this week', while ground truth is 'S.Korea talk with other country'. As you can find out from above example, two sentences are closely related but generated one is contains unnecessary information.
For time event extraction, dividing by person information seems reasonable but same problem of using HMM as mentioned above. Result of NER is also closely related to ground truth. For place extraction, it is 100% true. However, since it is Standford NER model, distinguishing Korean name shows error. For organization, it also tag not organization proper noun to organization frequently.

4.2 Issue Trend Analysis

Why LDA chose number of topics as 26?
Choosing topic number is based on evaluation method. We chose coherence values for metrics. Following figures are the relation between number of topics and coherence scores.

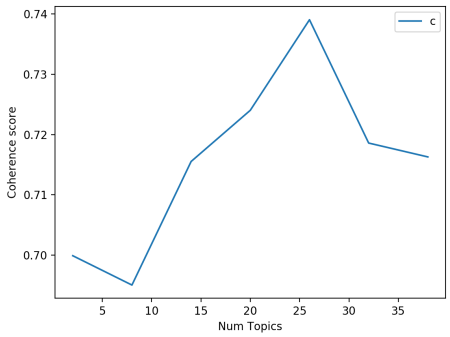

Figure 7: Coherence Score in number of topics (2015)

It is easy to see that coherence value is around 0.7 when number of topic is 26. We chose this value as number of topics and applied LDA. The graph for 2016, 2017 are in appendix. However, they show very low average coherence score (about 0.3~0.4). So we just selected number of topics as 26.
Why DBSCAN uses alternatives rather than Dunn Index for evaluation?
The formula for Dunn Index is as follows.

$$I_m = \frac{\min \delta(C_i, C_j)}{\max \Delta_k}$$

numerator is for distance between cluster and denominator for inter-cluster distance. Since we apply DBSCAN on every topic, it is not easy to calculate all distances (computationally inefficient). So, we chose alternative method. We plotted number of clusters as epsilon changes, and chose the epsilon with highest number of clusters. Let's say that epsilon value as E.
Reason for choosing epsilon with highest number of clusters is because as can see in figure 8, left part of peak point indicates there is only new forming cluster before epsilon becomes E. If new cluster formed, then minimum value of distance between cluster would decrease while maximum value of cluster diameter doesn't change. Therefore, as epsilon gets to value E, dunn index becomes better. However, right part of peak can't be explained only by epsilon value. It can vary as how points distributed over space. Therefore we chose epsilon value of peak point as optimal value.

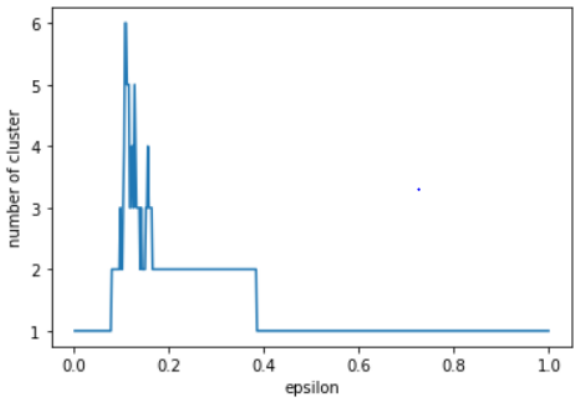

Figure 8: Coherence Score in number of topics (2015)

Why DBSCAN uses only documents title for clustering?
Because documents title represents the overall content in document. Also, DBSCAN is applied after topic modeling. So, documents are already topically related.

Issue Trend List Evaluation

Let issue set be $I = \{I_1, I_2, I_3, \ldots, I_{10}\}$. For issue I_k, Let' say issue extracted from our method is A_k, and issue from ground truth B_k.

N_k : the # of tokens(words) of A_k that are in B_k

M_k : the # of tokens(words) of B_k

$$score\ S = \frac{\sum_{i=1}^{10} N_i}{\sum_{i=1}^{10} M_i}$$

Year	Method A	Method B
2015	0.47	0.55
2016	0.39	0.44
2017	0.38	0.52

Figure 9: Score of Method A and Method B for each year

From above result, we could see that our trending list captures about half of information from the ground truth trending list.

4.3 On-issue event tracking

Why we choose NER method to extract time event?
By only looking at document title, it is hard to find out whether time continuous two document represents same event. Therefore, we thought two document representing same event should have same result of NER. So, we extracted person, location and organization of each document and cluster document which have same result of NER and continuous in time.

Why only consider person information while extracting time event?
Person is main agent of event. Therefore, person information can be considered as the most important information among NER results. Also, location and organization can be varied in same event. For example, in event 'U.S. defense chief visit', U.S. defense chief can visit many places around Korea, while most important fact is that U.S. defense chief visit Korea. Therefore, we only used person information while extracting time event.

How HMM works?

Therefore, we restricted length of sentence and chose central sentence from top 10 generated sentence. Reason for choosing central sentence from top 10 generated sentence is because generated sentence abstracts cluster and central sentence represent those abstracted sentences. Method of choosing central sentence is we vectorized the title from cluster and chose central vector. HMM through clustered document. We run HMM at data title of , and its result represent time event of corresponding issue, which is HMM result of . For HMM, we used bigram model which is simple but result is reasonable. From title of , we generated bigram model, and we generated random sentence. After generating random sentence, we calculated score of sentence in mean of conditional probability distribution of each word in sentence. We generated several random sentences and calculated score of each sentence and sorted in order. Sentence of highest score doesn't represent key phrase of cluster. Since score increase as sentence get shorter. Therefore, we restricted length of sentence and chose central sentence from top 10 generated sentence. Reason for choosing central sentence from top 10 generated sentence is because generated sentence abstracts cluster and central sentence represent those abstracted sentences. Method of choosing central sentence is we vectorized the title from cluster and chose central vector.

4.4 Related-issue event tracking

Why cluster in same topic can be said related but not directly related?
After LDA, those topics represents area rather than issue. In same topic, there can be many issue. We though biggest cluster issue represents top trending issue while other represents top trending issue's related event. Since it is in same topic it can be said related, but not in a same cluster represents not directly related. Therefore, cluster in same topic can be said related but not directly related.

NER Evaluation

Let event set be $E = \{E_1, E_2, E_3, \ldots, E_n\}$. For event E_k, Let' say named entity values extracted from our method is A_k, and entity values from ground truth B_k. (named entity can be PERSON, ORGANIZATION, PLACE)

- N_k: the # of entity values of A_k that are in B_k
- M_k: the # of entity values of the B_k
- R_k: the # of entity values of A_k that are not in B_k

$$score\ S = \frac{\sum_{i=1}^{n} N_i - R_i}{\sum_{i=1}^{n} M_i}$$

Since we have 5 events extracted from related-issue event tracking, we calculated the score with named entity values from those events. Result is as follows.

$N_k = \{18,20,17,21,20\}$
$M_k = \{21,24,22,25,27\}$
$R_k = \{5,3,4,2,3\}$

$$score\ S = \frac{\sum_{i=1}^{5} N_i - R_i}{\sum_{i=1}^{5} M_i} = 0.782$$

To be precise, we need more evaluations from other events. However, the score seems acceptable.

5 Work distribution

In this part, we will show work distribution throughout this project. Most of the part, we worked together, data preprocessing, DBSCAN and noise elimination. Junmo Cho focused on NER and evaluation. Byoungjun Kim focused on LDA and HMM.

REFERENCES

[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan, 2003. Latent Dirichlet Allocation. *J,Mach Learn. Res.*, 3:993-1022.

[2] Xiaohua Liu, Furu Wei, Shaodian Zhnag and Ming Zhou, 2013. Named entity recognition for tweets, *TIST volume 4 issue 1*.

[3] Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Workshop on Speech and Natural Language.* 112-116.

[4] Ping Yin, Ming Zhang, ZhiHong Deng and DongQing Yang, 2004. Metadata Extraction from Bibliographies Using Bigram HMM. *ICADL, International Collaboration and Cross-Fertilization,* 310-319

Appendix

2015 top 10 trending list

	Method A	Ground truth
1	S.Korea 's visit N.Korea to hold security dialogue	S.Korea talk with Japan, U.S., China about security about N.Korea nuclear
2	S.Korea, Poland agree to boost ties	S.Korea boost ties with other country
3	S.Korea, Japan to hold first security talks this week	S.Korea defense talk with other country
4	N.Korea appears to fire medium-range Nodong missile: U.S. expert	N.Korea missile test
5	Child care centers to bomb-making tips	S.Korea education phenomenon
6	Park urges parliamentary approval of labor next year	S.Korea labor
7	Seoul hails opening of U.N. resolution on N.K.	N.Korea human-right
8	S.Korea reports 2 more MERS, no new cases	MERS
9	S.Korea, U.S. to laumch joint drill	S.Korea, U.S. military drill
10	Former prime minister denies corruption	High ranking official corruption

Figure 1: Top 10 issue from method A and ground truth (2015)

	Method B	Ground truth
1	S.Korea 's visit N.Korea to hold security dialogue	S.Korea talk with Japan, U.S., China about security about N.Korea nuclear
2	Former Korean Air heiress faces further charge	Nut rage
3	Park urges parliamentary approval of labor next year	S.Korea labor
4	U.S. legislation calls for daycare center teacher	Daycare center problem
5	Seoul hails opening of U.N. resolution on N.K.	N.Korea human-right
6	Fewer Seoulities marry in 2014 as important : survey.	S.Korea social trends
7	S.Korea, U.S. to launch joint drill	S.Korea, U.S. military drill
8	Park's approval rating hits all-time low	Park's approval rating hits bottom
9	Child care centers to bomb-making tips	S.Korea education phenomenon
10	U.S. says he held talks with additional topic of arms control	U.S. and S.Korea talks about army control

Figure 2: Top 10 issue from method A and ground truth (2015)

2016 top 10 trending list

	Method A	Ground truth
1	S.Korea, U.S., Japan to hold phone talks	S.Korea hold talks with other counties
2	N.Korea officials involved in human rights foundation	N.Korea human rights
3	People's Party leaders of audit kicks off impeachment motion	President Park and Saenuri Party
4	N.Korea may deploy long-range rocket launchers	N.Korea missile
5	ex-presidential aides to be questioned over political scandal	Choi Soon-sil national affair scandal
6	Man gets life for murdering two teenage girls	Gangnam Station Murder
7	Number of child abuse	Social problems (poverty, divorce, wage)
8	N.Korea does not stipulate itself as nuclear capabilities ahead of reprocessing activity: 38 North	N.Korea nuclear
9	Korean teens fight for rights to instill loyalty	X
10	Korea, U.S. to hold talks on drones	X

Figure 3: Top 10 issue from method A and ground truth (2016)

	Method A	Ground truth
1	Number of child abuse	Social problems (poverty, divorce, wage)
2	N.Korea does not stipulate itself as nuclear capabilities ahead of reprocessing activity: 38 North	nuclear in N.Korea
3	N.Korea officials involved in human rights foundation	N.Korea human rights
4	People ' s Party leaders of audit kicks off impeachment motion	President Park and Saenuri Party
5	Defense Ministry to secure site for THAAD	THAAD
6	Korea confirms first Zika virus	Zika virus
7	Half of 10 young job market: data	job market
8	Korea set to unveil follow-up measures on sex slaves	comfort woman issue
9	ex-presidential aides to be questioned over political scandal	Choi Soon-sil national affair scandal
10	Man gets life for murdering two teenage girls	Gangnam Station Murder

Figure 4: Top 10 issue from method A and ground truth (2016)

2017 top 10 trending list

	Method A	Ground truth
1	Top diplomats of S.Korea, New Zealand discuss cooperation	N.Korea nuclear issue
2	Trump says there will take 'decisive': White House: Trump calls for stepping back to table in advance	Trump and N.Korea
3	Korea, US troops to upgrade Patriot missiles against N.Korea	S.Korea and U.S military drills
4	N.Korea likely to end in next 30 days: expert	N.Korea nuclear test
5	Special investigators ready to attend Park 's arrest	President Park Court issue
6	Presidential candidates having a field day online on Election Day after law change	presidential candidate
7	S.Korea mulling over how to accept Moon 's offer for PyeongChang Olympics	X
8	N.Korea 's grain imports from N.Korea trying to China dives in September	Trade between S.Korea and N.Korea
9	Police investigating online rape threat against child	Sexual crime issue
10	Court rejects arrest warrants for two ex-NIS officials	Rejection of ex-NIS in president Park Court

Figure 5: Top 10 issue from method A and ground truth (2017)

	Method A	Ground truth
1	Opposition parties urge Park ' s impeachment timeline	President Park impeachment
2	Elite female spy unit behind murder of late singer Kim 's half brother	Kim Jong-nam death
3	Top diplomats of S.Korea, New Zealand discuss cooperation	N.Korea nuclear issue
4	Trump says there will take 'decisive': White House: Trump calls for stepping back to table in advance	Trump and N.Korea
5	More seniors subject to Korea grows quickly over decade	Senior population growth
6	Korea to Japan 's action against 'comfort women deal	comfort women issue
7	Police investigating online rape threat against child	Sexual crime issue
8	Acting president orders prompt vaccination of foot-and-mouth disease	foot-and-mouth report
9	Moon vows all-out efforts to stamp out corruption in defense industry	president Moon jae-in
10	Presidential candidates having a field day online on Election Day after law change	presidential candidate

Figure 6: Top 10 issue from method B and ground truth (2017)

On-issue Tracking

```
Issue : S.Korea, Japan to hold first security talks this
week.
Event : Seoul, Tokyo to hold high-level economic talks,
S. Korea, U.S. to hold cyber-security talks, U.S. defense
chief to visit S. Korea: source, S. Korea, Japan to hold
series of defense, foreign officials' meetings, S.Korea,
Japan to hold first security talks in 5 yrs., S. Korea,
U.S. hold annual defense talks, S.Korea, Japan to hold
first defense ministers'talks., S. Korea, Japan to reopen
finance ministers' talks on end, S. Korea, Australia hold
working-level defense talks, Korea, Japan in talks to
resume working-level defense dialogue channel, S.Korea,
Japan to hold director-level defense talks., S. Korea,
Australia to hold joint talks of foreign, defense
ministers, S. Korea, U.S. to hold defense talks, Korea,
Japan defense chief to meet in Seoul, S. Korea, China to
hold talks on sea boundary


Detail information :


 Event : Seoul, Tokyo to hold high-level economic talks
 Person : Yasumasa Nagamine
 Organization :
 Place : Japan, Seoul, South Korea, Tokyo


 Event : S. Korea, U.S. to hold cyber-security talks
 Person :
 Organization : Sony Pictures
 Place : North Korea, Washington, United States, Seoul,
South Korea


 Event : U.S. defense chief to visit S. Korea: source
 Person : Kim Min-seok, Ashton Carter, John Kerry, Martin
Dempsey
 Organization : United States Joint Chiefs of Staff, Adm,
Terminal High-Altitude Area Defense, THAAD
 Place : North Korea, Alaska, Washington, Japan,
Northeast Asia, Afghanistan, Seoul, South Korea,
California, China, The United States


 Event : S. Korea, Japan to hold series of defense,
foreign officials' meetings
 Person :
 Organization : Korea-United States Integrated Defense
Dialogue, Defense Trilateral Talks
 Place : North Korea, Washington, United States, Japan,
Seoul, South Korea, China, Tokyo
```

Event : S.Korea, Japan to hold first security talks in 5 yrs.
Person : Shinzo Abe, Junichi Ihara, Akitaka, Lee Sang-deok, Lee Myung-bak, Tony Blinken
Organization : Oceanian Affairs Bureau, United States Undersecretary of State Wendy Sherman, United States-Japan Defense Cooperation Guidelines, United States-Japan Defense Cooperation, Yonhap News Agency
Place : North Korea, Washington, Japan, Earlier, Northeast Asia, United StatesSouth Korea, Seoul, China, Tokyo, Dokdo

Event : S. Korea, U.S. hold annual defense talks
Person : David, Ashton Carter, Yoo Jeh-seung
Organization : Korea-United States Integrated Defense Dialogue, Defense Trilateral Talks, United States-Japan defense, United States Defense
Place : North Korea, United States Forces Korea, Washington, Japan, Seoul, China

Event : S.Korea, Japan to hold first defense ministers'talks.
Person : Ashton Carter, Han Min-koo, Gen Nakatani
Organization : Acquisition, United States-Japan, United States Defense, Defense Ministry, Shangri-La Dialogue
Place : North Korea, Washington, United States, Japan, Australia, Seoul, South Korea, Tokyo, Singapore, Vietnam

Event : S. Korea, Japan to reopen finance ministers' talks on end
Person : Choi Kyung-hwan, Taro Aso
Organization : Toray, Taro, the Group of Twenty, Toray Industries
Place : Japan, Nagoya, Seoul, South Korea, Tokyo

Event : S. Korea, Australia hold working-level defense talks
Person : Scott Dewar
Organization : Director-General of International Policy Yoon Soon-ku, Korea-Australia defense, Defense Ministry
Place : United States, Asia-Pacific, Australia, Seoul, September

Event : Korea, Japan in talks to resume working-level defense dialogue channel
Person : Shinzo Abe
Organization :
Place : United States, Japan

Event : S.Korea, Japan to hold director-level defense talks.
Person : Gen Nakatani, Atsuo Suzuki, Yoon Soon-ku, Kim Min-seok, Zagdsuren Boldbaatar
Organization : Yonhap, General Security of Military Information Agreement, Defense Ministry
Place : North Korea, Mongolia, Japan, Uzbekistan, Seoul, Ulaanbaatar, Tokyo

Event : S. Korea, Australia to hold joint talks of foreign, defense ministers
Person : Voltaire Gazmin, Murray McCully, Yun Byung-se, Benigno, Kevin Andrews, John, Julie Bishop, Inoke Kubuabola
Organization : the Ministry of National Defense, Han, Defense Minister Han Min-koo
Place : Philippines, United States, Fiji, Australia, South Korea, New Zealand, Sydney

Event : S. Korea, U.S. to hold defense talks
Person : Abraham Denmark, Elaine Bunn
Organization : Korea-United States Integrated Defense Dialogue, OPCON, defense, the Ministry of National Defense, Defense Ministry
Place : North Korea, United States, East Asia, Seoul, South Korea

Event : Korea, Japan defense chief to meet in Seoul
Person : Han Min-koo, Gen Nakatani
Organization : Defense Ministry
Place : Japan, Seoul, United States

Event : S. Korea, China to hold talks on sea boundary
Person : Hua Chunying, Ieodo
Organization : ROK
Place : President Park, Beijing, Marado, BEIJING, Seoul, South Korea, East China Sea

Related-issue Tracking

Issue : S.Korea 's visit N.Korea to hold security dialogue.
Event : N.Korea fishermen back home., Koreas to hold high-level talks., U.S. denies asking President Park to attend China ceremony., N.Korea for inter-Korean exchanges., Chinese tours to lift Ebola restrictions: tour firm.

Detail information :

Event : N.Korea fishermen back home.
Person : surnamed Lee, Kim Jung-wook, Kim Kuk-gi, Shin Hyon-hee, surnamed Jin, Joo Won-moon
Organization : Yonhap, Red Cross, Unification Ministry, Three North Korean, Korean Central News Agency, New York University, Unification, Choe, Coast Guard, KCNA
Place : Pyongyang, Panmunjeom, Panmunjom, South Hamkyong Province, Ulleungdo, Seoul, China, East Sea, Wonsan, Panmumjom, United

Event : Koreas to hold high-level talks.
Person : Kim Jong-un, Jeong Joon-hee, Kim Yang-gon, Yoo Ho-yeol, Kim Ki-woong, Hwang Chol, Jon Jong-su, Hong Yong-pyo, Yang Moo-jin, Rodong Sinmun
Organization : Hong, Inter-Korean, Committee for Peaceful Unification, Committee, Inter-Korean Dialogue, United Front Department, United Nations Security Council, Unification Ministry, Kaesong Industrial Complex, They, the North, Breaking, Workers ' Party ' s United Front Department, North ' s Committee for Peaceful Unification, University of North Korean, `` The South, Committee for Peaceful Unification of the Fatherland, Unification Ministry ' s Special Office, Unification, Security, Peaceful Reunification of, Peaceful Reunification of Korea, Unification Minister Hong
Place : Pyongyang, Tongilgak, 2010, Mount Geumgangsan, Panmunjom, October, Seoul, But the North, The North, Mount Kumgang, Kaesong

Event : U.S. denies asking President Park to attend China ceremony.
Person : Moon Jae-in, Choe Ryong-hae, Vladimir Putin, Jiang Zemin, Guo Weimin, Zhang Ming, Japan, Hu Jintao, Tony Blair, Gerhard Schroeder, Kim Jong-un, Xi Jinping, Great Hall, James Hardy, Park, Kim Jong-il, Min Kyung-wook, Barack Obama
Organization : Hardy, Kyodo News, IHS Jane, Information Office of the State Council, Yonhap News Agency, Cheong Wa Dae, White House, the South Korean Foreign Ministry, Xinhua News, Gate of Heavenly Peace, Chinese Communist Party, Foreign Ministry, Asia-Pacific

Place : Beijing, Tiananmen Square, Egypt, Kazakhstan, South Africa, Myanmar, Pyongyang, North Korea, Seoul, Cambodia, Venezuela, Tajikistan, The United States, Mongolia, Shanghai, Russia, Japan, China, Laos, Washington, United States-South Korea, Xi Jinping

Event : N.Korea for inter-Korean exchanges.
Person : Marshal Kim Jong-un, Lim Byeong-cheol, Jeong Joon-hee, Min Kyung-wook
Organization : Yonhap, Unification Ministry, Kaesong Industrial Complex, The Ministry of Unification, Cheong Wa Dae, Saenuri Party, Korean Central News Agency, National Assembly, Korean Red Cross, The Unification Ministry, National Security Council, KCNA
Place : Pyongyang, North Korea, Panmunjom, Japan, Seoul, Washington, Chuseok, South Korea-United States

Event : Chinese tours to lift Ebola restrictions: tour firm.
Person :
Organization : National Tourism Administration, DPRK, Koryo Tours, Young Pioneer Tours
Place : Friday, West Africa, Dandong, Jilin, Pyongyang Marathon, North Korea, Liaoning, Sinuiju, Saturday, Monday, South Korea, Hunchun