News Stream Analysis

Dongyeun Lee
Korea Advanced Institute of Science and Technology
Daejon, Republic of Korea
ledoye@kaist.ac.kr

Sangwon Lee
Korea Advanced Institute of Science and Technology
Daejon, Republic of Korea
ddw02141@kaist.ac.kr

Sunwoo Im
Korea Advanced Institute of Science and Technology
Daejon, Republic of Korea
ism07@kaist.ac.kr

KEYWORDS

news analysis, event tracking, topic detection, named entity recognition

ACM Reference Format:
Dongyeun Lee, Sangwon Lee, and Sunwoo Im. 2019. News Stream Analysis. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The problem we need to solve is to find the most talked about issue in the news set and to find events associated with it. And the goal is to extract and print out their detailed information. The three problems were solved through the following ideas.

The first is the question of finding the most talked about issue. To solve this problem, the TF-IDF tried to rank the trends. However, when using TF-IDF, only one word is extracted and the meaning is hard to understand. And it was hard to categorize Korea as the importance of "Korea" was high in most of the documents. So we first proceeded with document clustering through LDA-Topic modeling. Later, we sought to find and select a group of the most duplicated words in the titles of classified documents. However, there was no way to know if the chosen title really represented the issue. Therefore, TF-IDF was calculated and selected the title of the document containing the high importance words as the title of the issue.

The second problem is the extraction of events related to the issue. There are two minor issues here: related-issue event tracking and on-issue event tracking. Among them, the related-issue event tracking related is expected to be resolved by one more clustering of documents clustered as issues. So we sorted out the documents clustered as one issue once more through LDA-Topic modeling. This time, we used cosine similarities to select the title of the event, the title of the document with the highest similarity to other documents within the event. This is because we thought that the number of documents per event would decrease the efficiency of the TF-IDF by classifying them twice, and that if there were many similar aspects to other documents, they could be representative of the event.

And we thought on-issue event tracking would be solved by just aligning the time sequence based on the related-issue tracking problem. However, once the documents were sorted in chronological order, they were determined not to be the sequence of events related to the issue but to be another on-issue event tracking related to the event. Therefore, we chose to arrange them in chronological order first and tie them together [4]. Using the cosine similarity, I thought that a high-altitude article reported within a short period of time would report the same event. After that, if these bundles were again calculated and had a high degree of similarity, they were arranged in chronological order to solve the problem.

Finally, the third problem was the problem of extracting detailed information. We used the CONLL2003 dataset to create AI, which learns IOB tagging as an LSTM-based model and automatically tags documents. NER tagging was done on the documents and the words that appeared most frequently for each event were presented as detailed information.

2 PROBLEM STATEMENT

2.1 Issue Trend Analysis

Issue trend analysis can be defined as

- From bunch of news articles, detect top trending issues for entire news article, and for each year / section
- Find top trending issues for news according to classification criteria will provide some insight from a large amount of information at a glance, even if you don't read all the news.

2.2 On-Issue Event Tracking

On-issue trend analysis can be defined as

- For some trending issues, describe events occurring in chronological order
- For each event, structuralize information as Person / Organization / Place by information extraction
- On-Issue Event tracking is expected to help following-up specific trendy issue

2.3 Realted-Issue Event Tracking

Related-issue trend analysis can be defined as

- Extract and describe related events that are not directly associated with a particular issue
- For each event, structuralize information as Person / Organization / Place by information extraction
- From the news provider's point of view, related-issue event tracking could help provide the following articles to recommend to

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

those who read specific article

3 RELATED WORK

3.1 TF-IDF

TF-IDF(Term Frequency–Inverse Document Frequency) is a method of calculating a significant degree of weight for every word in the Document-Term Matrix(DTM), using the frequency of the word and IDF. This allows you to compare documents more accurately than using DTM. Usually used to obtain similarity of a document or to obtain importance of a particular word within a document. To take advantage of this, it is necessary first to create DTM. And if the number of documents in which a particular word t appears is df(t), the idf(t) is calculated by the following formula. (idf(t) = log(n/ 1 + df(t)) n is the total number of documents and the reason for using log is because the larger the number of documents, the more exponentially the value of the idf. Also, the reason for adding 1 is to prevent the denominator from becoming zero when calculating words that do not appear in the document.

The TF-IDF determines that words commonly used in all documents are of low importance and words that frequently appear only in certain documents are of high importance. Smaller TF-IDF values are less important, and larger values are more important. The TF-IDF matrix, created for every word, is used in the LDAs to be described so that the importance of each word as well as the hidden meaning can be extracted.

3.2 LDA(Latent Dirichlet Allocation)

In text mining techniques is topic modeling of one of the algorithms. Assume that all documents are written as follows in order for the LDA to extract Topic from them. Determine the number of words to be used in the document and determine the mix of Topic based on probability distribution. And the words used in the document are selected based on the distribution of the probability of appearance of words in the Topic, which is statistically defined in the distribution of topics. And based on this assumption, the LDA performs reverse engineering, which backtracks the above process. If you tell me the number of topics, you will go through the process of assigning all words to the appropriate topic based on them. This process differs from one of the other Topic modeling algorithms, Late Semantic Analysis (LSA). While the LSA reduces the dimensions of the DTM or TF-IDF to group adjacent words into a single topic, the LDA calculates the topic as a statistical base, calculating the probability that the word used exists in the topic and the probability that the document contains topics. It is easy to find out how similar each Topic is and what topics are distributed by document through LDA

3.3 NER(Named Entity Recognition)

NER Tagging is about finding out what type of person, place, or group each word belongs to. There are various ways to do NER Tagging, but among them, We would like to introduce IOB Format. Tag each word with I, O, and B. I means the interior of an object, and O means the part that is not the object. B means the part where the object begins. It is usually tagged with B for what type it belongs to. CONLL2003 is a DATAset for NER tagging through the IOB method. (https://raw.githubusercontent.com/Franck-Dernoncourt/NeuroNER/master/neuroner/data/conll2003/en/train.txt)

4 PROJECT STRUCTURE

4.1 Issue Trend Analysis

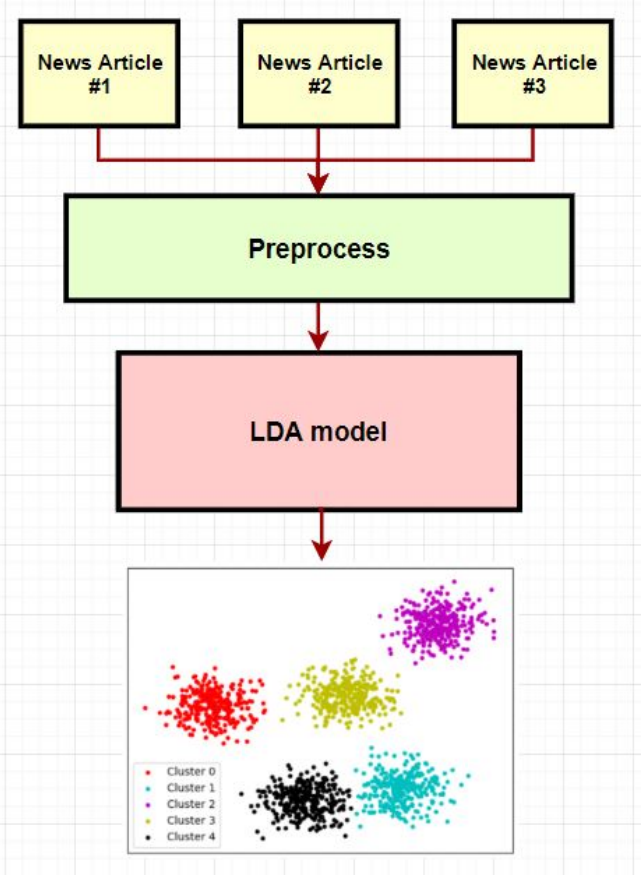

Figure 1: Issue trend analysis

For issue trend analysis, first we preprocess news data to meet our purpose. Second, using LDA model, we clustered articles with 125 issues and export issue tagged data for our consecutive works. News' body part is used for LDA model's input.

4.2 On-Issue Event Tracking

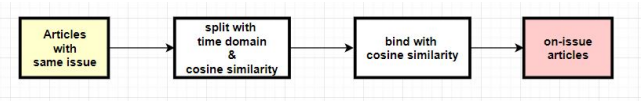

Figure 2: On-issue event tracking

For on-issue event tracking, first we receive issue tagged news article dataset from issue trend analysis. Then we make some set of articles depend on article's written time and cosine similarity

between articles. Then proceed similar step without time domain. We call the output of the second process as on-issue event set. On-issue event set contains sets which are connected in order of time on the same event.

4.3 Related-Issue Event Tracking

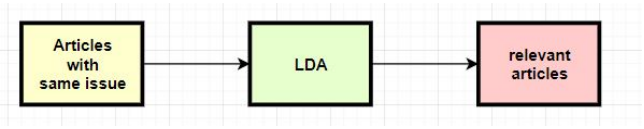

Figure 3: Related-issue event tracking

For related-issue event tracking, same as on-issue event tracking process, we receive issue tagged news article dataset from issue trend analysis. Then we make lda model for each issue. Each lda model proceed clustering on each issue and make 6 clusters.

4.4 NER Tagging

Figure 4: NER tagging

NER tagging is used in on-issue event tracking and realted-issue event tracking at the last part to extract information from event. From relevant articles, we use combined model of bi-LSTM and CRF to extract information of person, organizaiton, and location.

5 ISSUE TREND ANALYSIS

5.1 Issue Classification

Following steps are taken for a reasonable issue classification.

1. Remove meaning less words. We remove special characters such as ?, !, ", , < > and meaningless prefactory languages such as "[weekender]", "[newsmaker]", "[herald interview]".
2. Tokenize the words using nltk.wordtokenize library. Then lemmatize the words using nltk.stem.WordNetLemmatizer library.
3. Proceed LDA using gensim library. We set input of LDA model as number of cluster=125, iterations=400, passes=40. We will explain how we determine the number of clusters later.

5.2 Issue Naming

We ran issue naming using the title of the document. Through the LDA proceeded above, each document belongs to a specific issue. First we find the tf-idf of the words in the title belonging to the issue group. Then, for each group of issues, find the total number of consecutive phrases in the title. And only those statements with more than three occurrences of a particular statement in the issue group are listed. Three variables were set arbitrarily. Then, the sum of the tf-idf values of the words in each phrase in the candidate list is added. The candidate with the highest added tf-idf value becomes the issue name of the issue group.

6 ISSUE TRACKING

6.1 On-Issue Event Tracking

6.1.1 Making Event. For news that issues are classified through above process, we decided to focus more on time information. After deciding which issue to explore the event, we grouped the articles to meet the following two conditions for every article.

1. Articles within 7 days of the date of the first article
2. All articles in the group and articles with a cosine similarity of 0.5 or higher

We named such groups as "set". And we assume that each set represents events. To make this assumption concretely, one article was to belong to a maximum of one set.

6.1.2 Making On-Issue Event. Now measure the cosine similarity between the sets. For sets with cosine similarity 0.7 or higher, there are two options:

1. If the earliest time of the one set is earlier than the most recent article in the other set, combine two sets into one set.
2. If not, locate two sets "on-issue event set" without changing order.

The similarity between making a "set" is 0.5 and the similarity of 0.7 when making an "on-issue event set" is a practical reason. The analysis results showed that articles written at a relatively similar time often deal with the same event, even if the similarity is not too high. However, if the written times are far, the similarity should be larger than 0.5 to be judged as an article on the same issue.

6.2 Related-Issue Event Tracking

In the case of related issue event tracking, unlike on-issue event tracking, events do not need to occur in time order. So we did not consider the time when the news occurred. Of course, considering the time may yield better results, but we decided that the simple method we used produced good results. We selected an issue from the results of the issue trend analysis. After that, LDA was carried out in documents belonging to one issue, similar to issue trend analysis. In other words, two LDAs are used to obtain the results of related issue event tracking. Then, you can get the result through the event name selection and ner tagging process described below.

6.3 Event Name Selection

Event Name Selection algorithm is used in both on-issue event tracking and related-issue event tracking.

For making plausible event name, it was improper to use algorithm in issue trend analysis because the number of articles in a single set was relatively small. For this reason, we decide to pick representative article in a set. As for the method of selecting the representative article, it was conceived of who represented a group in our real life. When we think of a person who is a representative

of a group, we think of someone who can fit in with anyone in the group. Therefore, we thought that should be the case for articles representing one set. We pick the article with greatest average value of cosine similarity between whole articles in the set. From representative article, we extract the event name based on article's title and NER data.

6.4 NER Tagging

We used deep learning technology to process ner tagging. The model used bi-LSTM + Conditional Random Field (CRF). The reason why we used this was that we wanted to include deep learning technology in the project structure and decided that the best place to go was ner tagging. The Bidirectional LSTM CRF model is the most representative model used in the ner tagging process. The training data that we used was Annotated Corpus for Named Entity Registration. We used 80% of these data as training data and 20% as evaluation data.

CRF.PNG

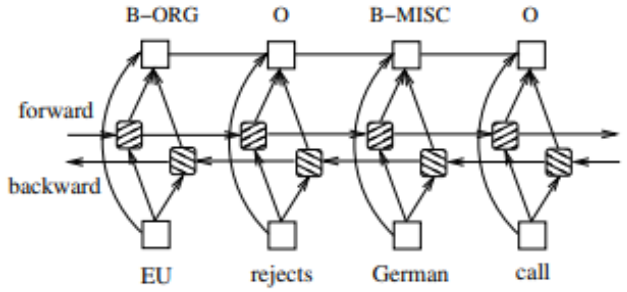

Figure 5: A BI-LSTM-CRF model. The picture is from the paper Bidirectional LSTM-CRF Models for Sequence Tagging[2]

Now let me explain how the training model was applied to our project. First of all, of course, we worked to match our data set to the input of the model. Then, all the results were saved. There are several sentences in each document, from which ner data was extracted and stored. Instead of saving one person, organization, and place for each document, all of the extracted entities are stored separately by tag. This data is located in the 'data/ner_tagged_news' directory. We use this data to proceed 'On-Issue Event Tracking' and 'Related-Issue Event Tracking'. Then, if there is an event set for ner tagging, extract the ner data of the documents included in the event set that you saved in advance. Then, extract the most frequently produced entity for each person, organization and place entity and export it to the result.

7 RESULTS

7.1 Issue Trend Analysis

Table 1 shows top 10 trend issue among given dataset.

7.2 On-Issue Event Tracking

Table 2 shows on-issue event tracking data from the issue#2 "Sanction (N.) Korea".

Rank	Issue
1	Ruling Opposition Party
2	Sanction (N.) Korea
3	Park Impeachment
4	Korea Urge (N.) Korea
5	Korea Japan China Hold
6	Visit Korea Discuss
7	Korea Strives
8	Korea Japan Slavery
9	(N.) Korea Missile
10	(N.) Korea Leader

Table 1: Result of Issue Trend Analysis

Date / NER components	Event
2015-04-02	Pelosi urges Abe to apologize over wartime sex slavery
Person / Organization / Place	Abe / Congress / Japan
2015-04-29	Abe refuses again to apologize for wartime sexual slavery
Person / Organization / Place	Abe / Congress / Japan
2015-07-29	U.S. lawmakers urge Abe to offer sincere apology for sexual slavery
Person / Organization / Place	Abe / Congress / Japan
2015-08-27	S. Korean students urge Abe to apologize for sexual slavery
Person / Organization / Place	Abe / Foreign Ministry / Japan
2015-11-24	Lawmakers from 5 countries launch anti-sexual slavery coalition
Person / Organization / Place	Kelly / New Zealand Fiona Claire Bruce / Japan

Table 2: Result of On-Issue Event Tracking on Korea Japan Slavery issue

7.3 Relate-Issue Event Tracking

Table 3 shows related-issue event tracking data from the issue#1 "Ruling Opposition Party".

The higher the quality, the more uniform and the lower the quality, the more likely the other topics will be mixed per topic. However, a high degree of coherence is not necessarily good. The higher coherence, the more complete the document must be the same, for it to be classified as the same topic. In comparison, the lower the Perplexity, the better the performance of the Topic model. Looking at the graph, both coherence and perplexity are decreasing. We saw perplexity as a more important factor. Thus, 125 topics that are no longer reduced in perplexity were chosen as the number of topics

Similarly, related-issue event tracking using Topic modeling used the same method to calculate the coherence and perplexity values according to the number of topics. To apply topical modeling once more to the issue clustered documents, the number of topicals was calculated to be less than 10 as there were fewer documents

Event#1	Parties push for family reunion with NK sparks controversy
Person / Organization / Place	Rep / National Assembly / Seoul
Event#2	Lawmakers raised W53.5B in donations last year
Person / Organization / Place	Rep / Justice Party / Busan
Event#3	Moon to meet ruling party officials over N. Korea, bipartisan efforts
Person / Organization / Place	Rep / National Assembly / Korea
Event#4	Ruling party reeling from election rout
Person / Organization / Place	Rep / Saenuri Party / Seoul
Event#5	Opposition presses Saenuri for P.M. removal
Person / Organization / Place	Lee / Saenuri Party / Korea
Event#6	Saenuri boycotts Assembly audit
Person / Organization / Place	Lee / Saenuri Party / Korea

Table 3: Result of Related-Issue Event Tracking. We select Ruling Opposition Party issue

per issue. As a result, the value of the coherence is high and the perplexity is low at four topics

8 EXPERIMENTS

8.1 Topic Modeling

What is important in Topic modeling is "Number of topics". We measured the coherence and perplexity of the model to evaluate the Topic model. The graph of the results measured with varying number of topics is shown in figure below.

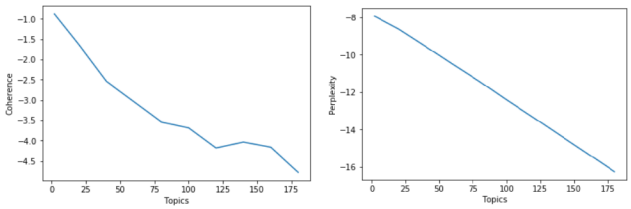

Figure 6: Coherence and Perplexity vs number of topics in Issue Trend Analysis

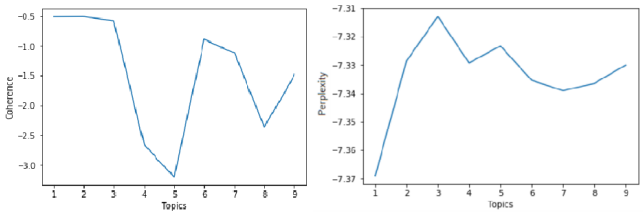

Figure 7: Coherence and Perplexity vs number of topics in Related-Issue Event Tracking

The higher Coherence is the better. But, a high degree of coherence is not necessarily good. Because if the higher Coherence, the more complete and identical documents are classified in the same topic. In comparison, the lower Perplexity is better performance of the Topic model. [5][1] Looking at the graph, both coherence and perplexity are decreasing. perplexity is almost linear. So it is difficult to decided topic number. We saw perplexity as a more important factor than coherence. Thus, 125 topics that are no longer reduced in coherence.

Similarly, related-issue event tracking using Topic modeling used the same method to calculate the coherence and perplexity values according to the number of topics. To apply topical modeling once more to the issue clustered documents, the number of topics was calculated to be less than 10 as there were fewer documents per issue. The result was set to six with high value of coherence and low decrease of perplexity.

8.2 NER Tagging

Entity	Precision	Recall	f1-score
Org	0.68	0.56	0.61
Per	0.80	0.75	0.77
Gpe	0.96	0.94	0.95
Geo	0.81	0.87	0.84
Tim	0.85	0.85	0.85
micro avg	0.82	0.79	0.81
macro avg	0.81	0.79	0.80

Table 4: Evaluation Result of Annotated Corpus for Named Entity Recognition Data

Entity	Precision	Recall	f1-score
Org	0.25	0.19	0.22
Per	0.45	0.35	0.40
Gpe	0.53	0.69	0.40
Geo	0.49	0.51	0.50
Tim	0.18	0.62	0.28
micro avg	0.35	0.42	0.38
macro avg	0.39	0.42	0.39

Table 5: Evaluation Result of Korea Herald News Data

We trained the Bidirectional LSTM CRF model with the 'Annotated Corpus for Named Entity Registration' data set. 80% of the data was used for training and 20% for evaluation. In addition, we directly ner tagging 130 of the news data sets we received. Because the data set that we trained and the data that we intended are different, we decided that performance evaluation through our data set should be carried out in order to perform a direct performance. The result was distinctly different. You can see the results by looking at Table 4 and Table 5. If you look at the evaluation result of the data set that was originally trained, you can see that f1-score is 0.80 and if you look at the evaluation result of the data set that we tagged

ourselves, you can see that f1-score is 0.38. It was a natural result, and although some difference was expected, there were many differences in performance. Since our data set is a translation of Korean news articles, unlike the training dataset, most NER data are people, place and organizations in Korea. Therefore, it is not easy to appear in the original training data. According to Table 5, the results are good in order of geopolitical entity, geographic identity, person and organization. Geopolitical entity contains words like Korean, Russian, and Japanese, and in this case the original data set also contains. That's why the results are the best. The same is true of 'geographic identity'. Conversely, in the case of 'organization,' our data set is Korean news, so it contains a lot of organization in Korea, which is almost hard to find in the set of data for training purposes. So the result was a f1-score of 0.22, which was hard to get a good result.

9 CONCLUSION

9.1 Issue Trend Analysis

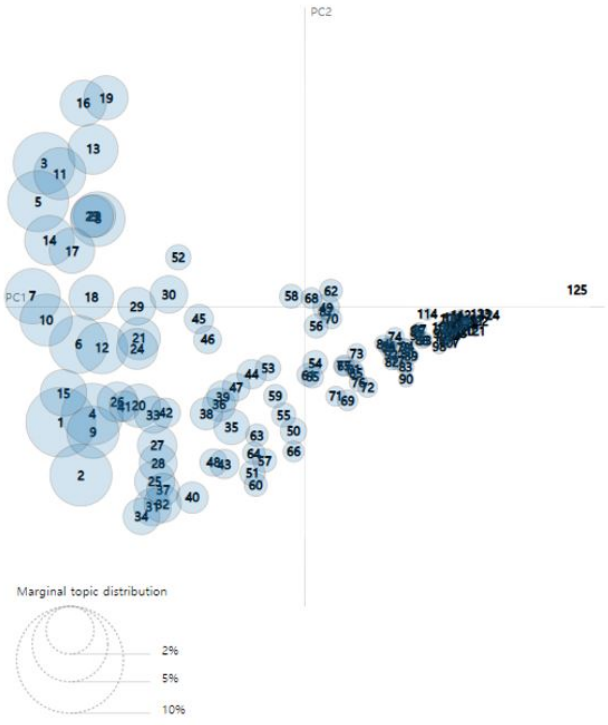

Figure 8: Issue trend analysis clustering result

The issue trend analysis result[Figure 8] shows that the documents are clustered in a total of 125 topics. However, There are many overlapping parts because the same word is in the upper ranking of several clusters. In addition, some topic does not have any documents. And if you look at the ranking of issue trends, you will find many overlapping names. I thought these problems were because many documents used similar word pools. In particular, we concluded that the title and cluster seem to overlap, as the word, especially "Korea", appears a lot in almost every document.

Consequently, the Topic modelling based on LDA model allowed the clustering of news sets to be classified and ranked according to the number of documents in the classified cluster. But there are still questions about whether the title of the topic was extracted properly. As duplicate titles have been extracted, it can be seen that the solution to the problem is more uncertain. There was no indication that the extracted title could represent the issue or event properly. It labels issues or events for all documents. And predicts that the word2vec model and use mean shift algorithm will help better performance than TF-IDF model.

9.2 On-Issue Event Tracking

Most of the cases, on-issue event tracking seems work properly. However, sometimes there are exceptions.

	Date	Event
Article1	2015-05-26	S. Korea refers N. Korea to U.N. sanctions panel over SLBM
Article2	2015-09-17	China may have offered 500,000 tons of crude oil to NK this year: Seoul

Table 6: Exception case of on-issue event tracking

In "Sanction (N.) Korea" issue, it was confirmed that the above two unrelated sentences were included in one on-issue event and we decide to observe the two articles in a more detail. Observations are like below.

	Cosine Similarity	Frequent words (in order)
Article1	0.42	[korea, north, sanction, committee]
Article2		[north, china, korea, tons]

Table 7: Comparison of two exception articles

This result is interpreted as a conflict between event name selection and making on-issue event. Because on-issue events use the average value of cosine similarity in each set, there could be cases where cosine similarity is less than 0.7 when one news article is drawn from each set and compared.

Then we can consider to select a representative article that is most similar to the previous set, which is as if it were a case of a main passenger transfer where the result is aligned and the algorithm is applied because the representative article does not represent the set.

Therefore, to solve this problem, it might be helpful to 1) change the metric of grouping on-issue event sets instead of cosine similarity (ex. Euclidean resistance, Jaccard similarity) 2) Instead of CounterVectorizer, whcih represents the word as combination of one-hot vectors, use word embedding to express the relationship between languages on a n-dimensional plane, such as Fast2Text and Glove.

9.3 Related-Issue Event Tracking

Extraction of events seemed to overlap a lot in the results [Figure 9], so they were poorly clustered. However, if you look at the percentage of words by topic [Figure 10], you can see a big difference.

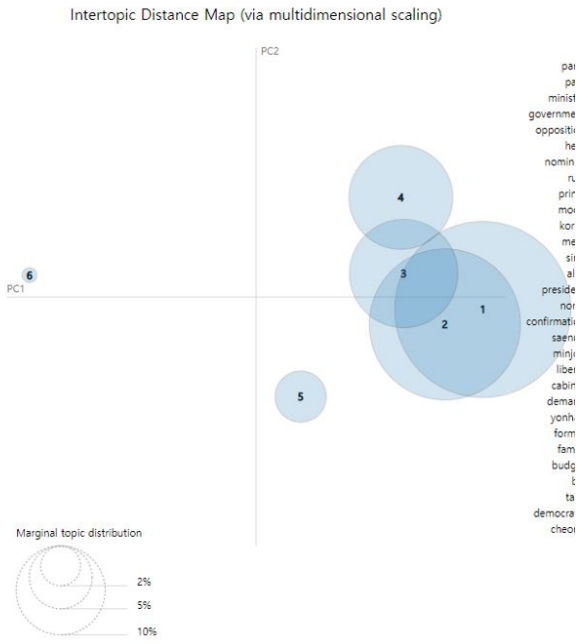

Figure 9: Related-issue clustering result

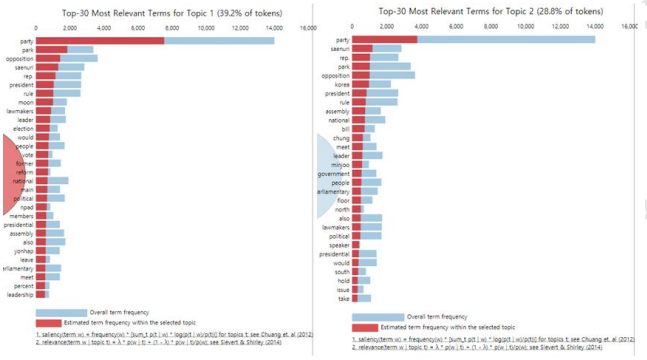

Figure 10: Percentage of words by topic

This is thought to have caused more overlap because the pool of words is so narrow that it has already attempted to clustered them on the same issue. However, given the weight of the words, there is a big difference, so the event was well classified. In other words, related-issue events tracking was successful.

Issue ranking and related-issue event tracking are based on the LDA model. Since LDAs are statistical models, they are not using the Neural network, so they could perform better. Indeed [3] studies show that models combine LSTM and LDA have better predictive performance than those using only LDA models. Although different from the objective of this project, it is expected that the application will produce equally good results. A good performance of Topic modeling will improve the ability to extract issues or events from a set of news.

9.4 NER Tagging

As I said above, it's a bit disappointing. The reason is that there is a large difference in the data set. For example, in the case of President Park Geun-hye, the news articles were often abbreviated as Park. In this case, our model was often interpreted as 'park', meaning park, and judged as a geographical entity. If our topic was not about Korean news but about American news, we could expect much better results. Also, if there were three data for ner tagging on Korean news, good results would surely be achieved.

ACKNOWLEDGMENTS

10 CONTRIBUTION

We worked together to discuss the structure of whole the project. The bottom part is the implementation part of each member.

10.0.1 DongYeun Lee. Project Structure Design
NER tagging - Whole part.
Issue Name - whole part except make_text_topic_list function
Suggest idea of every evaluation.
lemma_NER and extractNER function in on_issue_event_tracking.py and related_issue_event_tracking.py
IOB Tagging for NER evaluation

10.0.2 SangWon Lee. Followings are my contribution on this project.
Issue Trend Analysis - Preprocess steps and lda model codes
Issue Name - make_text_topic_list function - extracting most overlapped title which sequence length is 1 to 8 for each issue.
On-Issue Event Tracking - Almost every part
IOB Tagging for NER evaluation

10.0.3 SunWoo Im. Flow of problems solve implementation design
Issue Trend Analysis – lda model evaluation and lda model parameter tuning
Related-Issue Event Tracking – model design, model implementation, lda model evaluation and lda model parameter tuning
IOB Tagging for NER evaluation

REFERENCES

[1] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 288–296. http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf

[2] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[3] Yohan Jo, Lisa Lee, and Shruti Palaskar. 2017. Combining LSTM and Latent Topic Modeling for Mortality Prediction. arXiv:cs.CL/1709.02842

[4] Ida Mele and Fabio Crestani. 2017. Event Detection for Heterogeneous News Streams. In *Natural Language Processing and Information Systems*, Flavius Frasincar, Ashwin Ittoo, Le Minh Nguyen, and Elisabeth Métais (Eds.). Springer International Publishing, Cham, 110–123.

[5] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 100–108. http://dl.acm.org/citation.cfm?id=1857999.1858011