

# E2A

Extract 요약물 기반으로 한  
Abstract 요약 생성

- 이수빈, 이인재, 조양규, 최서현, 최희정 -

## 목차



# 1. 주제 선정 배경

# 1. 주제 선정 배경

- 사람은 어떻게 요약을 할까?



< 중요한 문장 선별 >

[헤럴드경제=신대원 기자] 박근혜 대통령은 새해 첫날인 1일 청와대 상춘재에서 출입기자단과 신년인사회를 겸한 티타임을 갖고 최순실 국정농단 파문 등에 대한 질문을 주고받았다.

박 대통령은 먼저 "국민들께 미안한 생각으로 무거운 마음으로 지내고 있다"면서 "하루 빨리 안정을 찾아 나라가 발전의 탄력을 키워야 한다는 마음"이라며 국회 탄핵 가결 이후 직무정지된 상황에 대한 심경을 토로했다.

< 최종 요약 >

朴대통령 "국민께 미안한 생각... 무거운 마음으로 지내"

출처: <http://news.heraldcorp.com/view.php?ud=20170101000208>

# 1. 주제 선정 배경

---

- Extract 요약 vs Abstract 요약

Extract 요약	Abstract 요약
문서에서 구나 문장을 그대로 추출하는 요약	문서의 내용을 압축하여 새로운 문서를 만드는 요약
요약문의 응집도나 가독성 확보가 어려운 단점 존재	자연어의 이해와 생성 기술이 필요하다는 어려움 존재
TextRank, LexRank	Seq2Seq

## 2. model 설명

# 1. 주제 선정 배경

- 사람은 어떻게 요약을 할까?



< 중요한 문장 선별 >

[헤럴드경제=신대원 기자] 박근혜 대통령은 새해 첫날인 1일 청와대 상춘재에서 출입기자단과 신년인사회를 겸한 티타임을 갖고 최순실 국정농단 파문 등에 대한 질문을 주고받았다.

박 대통령은 먼저 "국민들께 미안한 생각으로 무거운 마음으로 지내고 있다"면서 "하루 빨리 안정을 찾아 나라가 발전의 탄력을 키워야 한다는 마음"이라며 국회 탄핵 가결 이후 직무정지된 상황에 대한 심경을 토로했다.

< 최종 요약 >

朴 대통령 최순실 사태 마음 얘기

## 2. Model 설명

---

### 1. Extract 요약 알고리즘

: weighted-graph를 기반으로 문서에서 구나 문장을 그대로 추출하는 요약 알고리즘

#### 1) TextRank

: 문장을 벡터로 표현하여 쿼리와 문장들 간의 유사도를 판별하고, 이를 바탕으로 순위를 매겨 중요한 문장을 추출하는 요약 알고리즘 -> [gensim의 summarization](#) 이용

#### 2) LexRank

: 두 문장 사이의 유사도 척도로 IDF-modified cosine을 사용하고, 이를 바탕으로 순위를 매겨 중요한 문장을 추출하는 요약 알고리즘 -> [lexrankr의 LexRank](#) 이용

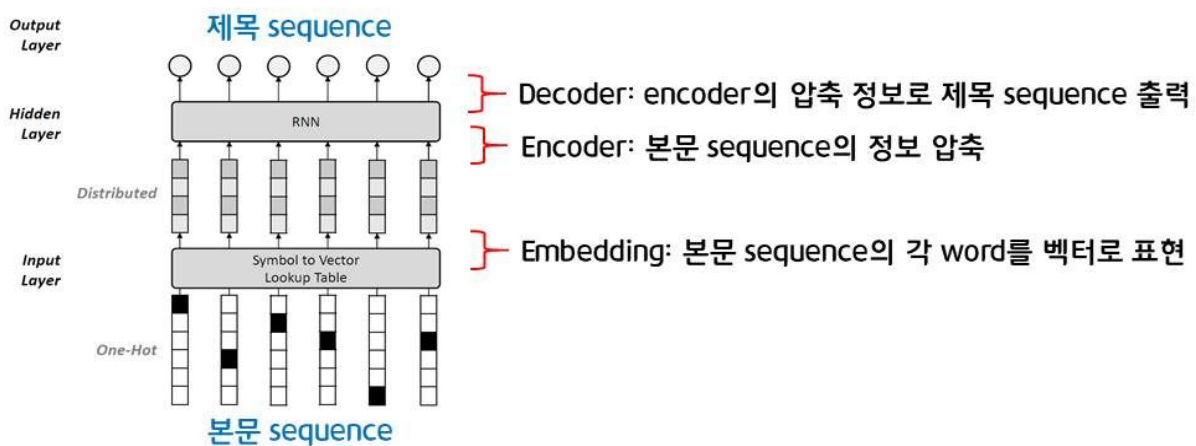


## 2. Model 설명

### 2. Abstract 요약 알고리즘

#### 2-1. Sequence-to-Sequence

: Seq2Seq는 RNN의 가장 발전된 형태의 모델로 LSTM, GRU 등 RNN cell을 길고 깊게 쌓아 sequence 데이터를 처리하는 데 특화된 모델

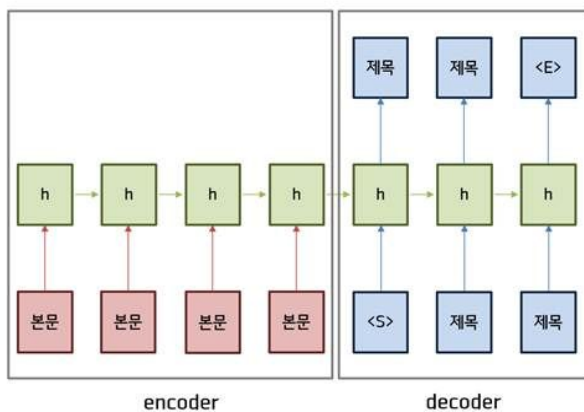


## 2. Model 설명

### 2. Abstract 요약 알고리즘

#### 2-1. Sequence-to-Sequence

##### - Encoder & Decoder



- Encoder input  
: 지정된 길이의 본문 sequence
- Decoder input:  
train: <S> + 지정된 길이의 제목 sequence  
test: <S> (이후 input은 예측된 단어)
- Target  
: 지정된 길이의 제목 sequence + <E>

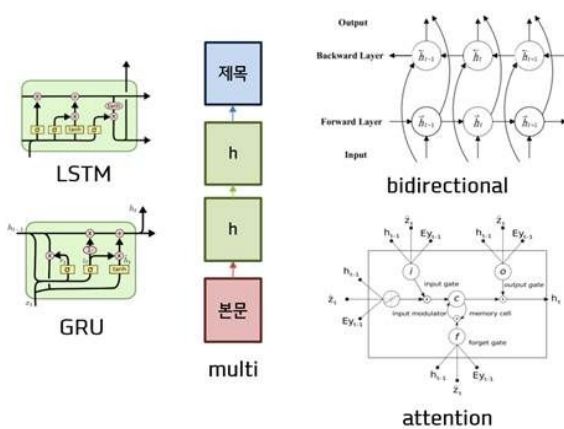
출처: <https://ratsgo.github.io/natural%20language%20processing/2017/03/12/s2s/>

## 2. Model 설명

### 2. Abstract 요약 알고리즘

#### 2-1. Sequence-to-Sequence

- seq2seq의 hidden layer 구성



- 원하는 RNN cell과 layer 수로 hidden layer 구성
- 2 layer의 BiLSTM으로 hidden layer를 구성하고 decoder에서 attention mechanism 이용

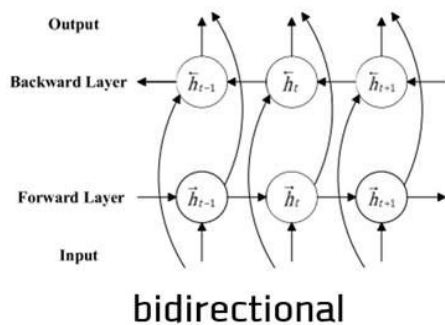
## 2. Model 설명

### 2. Abstract 요약 알고리즘

#### 2-2. BiLSTM with attention

##### - BiLSTM

: forward & backward 방향의 input을 모두 고려하는 LSTM



- Forward Layer  
: input sequence를 순방향으로 넣어 이전 정보를 저장
  - Backward Layer  
: input sequence를 역방향으로 넣어 이후 정보를 저장
  - Output  
: Forward Layer & Backward Layer를 모두 반영
- > BiLSTM은 이전 정보와 이후 정보를 모두 반영하므로  
이전 정보만 반영하는 기존 LSTM보다 성능 향상

출처: <https://ratsgo.github.io/natural%20language%20processing/2017/03/12/s2s/>

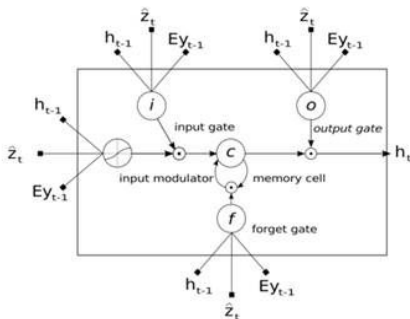
## 2. Model 설명

### 2. Abstract 요약 알고리즘

#### 2-2. BiLSTM with attention

##### - Attention mechanism

: Attentional decoder는 decoding시 매 time-step 별로 새로 생성될 토큰을 결정할 때 source sequence 중 가장 관련이 많은 token을 결정한 후 이 정보를 활용하는 구조



attention

- 문장 길이가 길고 층이 깊으면, encoder가 압축해야 할 정보가 너무 많아서 정보 손실이 일어나고, decoder는 encoder가 압축한 정보를 초반 예측에만 사용하는 경향을 보임
- 따라서 encoder-decoder 사이에 bottle-neck 문제가 발생함
- 이를 보완하기 위해 decoder 예측 시 가장 의미 있는 encoder 입력에 주목하게 만드는 attention mechanism 이용

출처: <https://ratsgo.github.io/natural%20language%20processing/2017/03/12/s2s/>

### 3. Modeling 과정

### 3. Modeling 과정

#### 1. 데이터 전처리

##### - Data set

article_date	article_press	article_url	article_title	article_cont
2017-01-01 10:17	아시아경제	http://news.naver.com	‘삼성’ 합병 노와주라 지시한 적 없다* 핵심	박근혜 대통령이 1일 오후 청와대 상춘재에서
2017-01-01 10:13	머니투데이	http://news.naver.com	박근혜 대통령 1일 새해 간담회 전문	글쎄 선택 본문 텍스트 작게 본문 텍스트 크
2017-01-01 10:10	한국경제	http://news.naver.com	2017년 신년 일상은 “국민들께 미안”...달라 ‘권한정치’ 아니던 2014-2016년에는 신년사	
2017-01-01 10:04	머니투데이	http://news.naver.com	“고조 겪어 마음 아파”...朴 측근·재벌에 등장 [CBS노컷뉴스 장관순 기자]1일 예정에 없던	
2017-01-01 9:45	이데일리	http://news.naver.com	朴대통령, 각종 의혹 전면 부인...현재-특검 ! [서울=뉴시스]전신 기자 = 박근혜 대통령이	
2017-01-01 9:41	이데일리	http://news.naver.com	[종합]朴대통령 전격 의욕 전면부인...현재, 박근혜 대통령이 정유년 새해 첫날인 1일 오	
2017-01-01 9:26	서울경제	http://news.naver.com	朴대통령 “물렀다. 었은것” 맞불...특검 지열 박 대통령, 기자단 신년 인사회 (서울=연합뉴	
2017-01-01 9:22	서울경제	http://news.naver.com	특검, 박근혜 정부 최순실 예산 내역 조사... 출근하는 박영수 특검 (서울=연합뉴스) 이경	
2017-01-01 9:15	아시아경제	http://news.naver.com	특검, 삼성 ‘부정정탁’ 수사 본격화...이변주 특검, 삼성 수뇌부 이변 주 출소환 (서울=연	
2017-01-01 9:15	머니투데이	http://news.naver.com	23일만에 침묵 갠朴대통령, 심경 토로 의욕 박근혜 대통령이 1일 청와대 상춘재에서 출	
2017-01-01 9:15	머니투데이	http://news.naver.com	朴대통령, 새해 첫날 깜짝 기자간담회...무엇 특검·언론의 불리한 여론 조성에 불만 표출	
2017-01-01 9:15	머니투데이	http://news.naver.com	朴 “나를 완전히 었은 것”...최순실·뇌물죄 도 긴급 기자간담회 열어 조목조목 반박 “너무나	
2017-01-01 9:14	머니투데이	http://news.naver.com	朴대통령 “뇌물죄, 완전히 었은 것...세월호대 박 대통령, 기자단 신년 인사회 (서울=연합뉴	
2017-01-01 9:13	이데일리	http://news.naver.com	朴, 삼성 합병 지시 혐의에 “완전히 었은것” [앵커]박근혜 대통령은 또 특검 수사와 현재	
2017-01-01 9:13	서울경제	http://news.naver.com	특검 “정유라 학정 특혜” 소설가 이인화 구4 특검 소환조사 받는 류철균 이대 교수 (서울	
2017-01-01 9:11	이데일리	http://news.naver.com	朴 “대통령도 사적 영역 있다”...간담회 성시 [앵커]박근혜 대통령은 간담회 내내 착잡한	
2017-01-01 9:10	이데일리	http://news.naver.com	朴 대통령 “세월호 구조에 신경 집중...다른 [앵커]박근혜 대통령이 정유년 신년을 맞아	
2017-01-01 9:06	이데일리	http://news.naver.com	[종합]국민의당 “朴, 피해자인양 위선 떠는 [서울=뉴시스]전신 기자 = 박근혜 대통령이	
2017-01-01 9:05	서울경제	http://news.naver.com	朴 대통령, 혐의 전면 부인...특검, 물증·전술 박근혜 대통령이 정유년 새해 첫날인 1일 오	
2017-01-01 9:00	아시아경제	http://news.naver.com	정의당 “朴, 국민을 바보로 아나” [서울=뉴시스]전신 기자 = 박근혜 대통령이	
2017-01-01 6:36	머니투데이	http://news.naver.com	朴대통령 “세월호때 할 것 다했다”... 현재 문박근혜 대통령이 정유년 새해 첫날인 1일 오	
2017-01-01 6:35	머니투데이	http://news.naver.com	朴대통령 “주사제는 사적 영역...국가순애 문박근혜 대통령이 1일 오후 청와대 상춘재에서	
2017-01-01 6:32	머니투데이	http://news.naver.com	박 대통령, “상춘재에서 아버지와 오찬 나누 박근혜 대통령이 1일 청와대 상춘재에서 출	

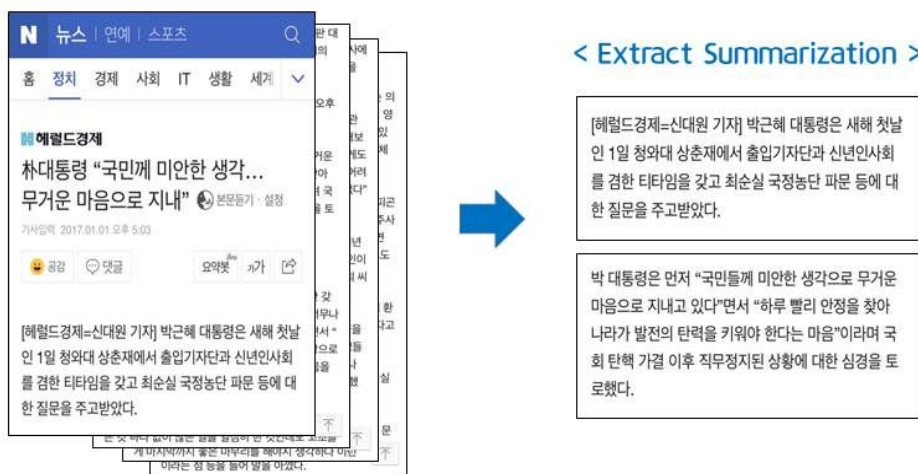
- 2017년 정치분야 기사(약 19000개) 본문 및 제목 크롤링
- 기사의 제목은 기사의 내용을 요약한 Abstract Summarization이므로 뉴스 기사를 데이터로 선정
- requests 모듈의 put 함수를 이용해 띄어쓰기 및 맞춤법 검사 진행
- 특수문자 및 영어, 숫자 제거

# 3. Modeling 과정

## 1. 데이터 전처리

### 1-1. Extract Summarization

: LexRank & TextRank 알고리즘을 기사 본문에 적용해 **extract summarization 데이터** 생성



출처: <http://news.heraldcorp.com/view.php?ud=20170101000208>



### 3. Modeling 과정

#### 1. 데이터 전처리

##### 1-2. word tokenize & normalization

###### - normalization

- : 사전에 구축한 stopwords 사전을 이용해 단어를 1차 normalize하고,  
빈도수가 1인 단어들에 대해 앞 2글자로 단어를 변경해 2차 normalize 진행  
(단, summary의 경우 stopwords 생성을 위해 stopwords를 제거하지 않고 구분하는 방식으로 진행)

###### < tokenize >

[ 박, 대통령은, 먼저, 국민들께, 미안한, 생각으로,  
무거운, 마음으로, 지내고, 있다면서, 하루, 빨리,  
안정을, 찾아, 나라가, 발전의, 탄력을, 키워야,  
한다는, 마음이라며, 국회, 탄핵, 가결, 이후,  
직무정지된, 상황에, 대한, 심경을, 토로했다 ]

[朴, 대통령, 국민께, 미안한, 생각,  
무거운, 마음으로, 지내 ]

###### < 1차 normalization >

[ 박, 대통령, 먼저, 국민, 미안한, 생각,  
무거운, 마음, 지내고, 하루, 빨리,  
안정, 찾아, 나라, 발전, 탄력, 키워야,  
마음, 국회, 탄핵, 가결, 이후,  
직무정지, 상황, 대한, 심경, 토로 ]

[朴, 대통령, 국민, 께, 미안한, 생각,  
무거운, 마음, 으로, 지내 ]

###### < 2차 normalization >

[ 박, 대통령, 먼저, 국민, 미안한, 생각,  
무거운, 마음, 지내, 하루, 빨리,  
안정, 찾아, 나라, 발전, 탄력, 키워,  
마음, 국회, 탄핵, 가결, 이후,  
직무정지, 상황, 대한, 심경, 토로 ]

[朴, 대통령, 국민, 께, 미안한, 생각,  
무거운, 마음, 으로, 지내 ]

### 3. Modeling 과정

#### 1. 데이터 전처리

##### 1-3. Text & Summary 유사도 상위 70% 데이터 선정

: Text의 문서 벡터와 Summary의 문서 벡터를 이용해 Text & Summary간 유사도를 계산하고, 이를 기준으로 상위 70% 유사도 데이터(약 14000개) 선정

abstract_sum	extract_sum	sum_type	sim
삼성 합병 도와주라 지시한 적 없다 핵심 청와대 제공 박근혜 대통령 일 오후 청와대		textrank	114.8096161
년 신년 일성은 국민 들께 미안 달라진 박 권한정지 아니 년 신년사 통해 국정 자신감		lexrank	94.37200928
고초 겪어 마음 아파 박 측근 재벌 에 동정 박 대통령 기업인들 생각 거기 미안한 마음		textrank	97.27378845
박 대통령 각종 의혹 전면 부인 현재 특검 상황 박 대통령 현재 심판 특검 수사 적극적		lexrank	107.3800354
박 대통령 전격 의혹 전면 부인 현재심판 (박 대통령 자리 세월호 참사 당일 미용 시술		lexrank	139.0446777
박 대통령 몰랐 다 엮은 것 맞불 특검 치박 대통령 새해 첫날인 이날 청와대 상춘재		textrank	90.59164429
특검 박근혜 정부 최순실 예산 내역 조사 이보배 박근혜 최순실 게이트 파헤치 박영		lexrank	129.9519653
특검 삼성 부정청탁 수사 본격화 이번 주 최지성 박상진장충기 거론 채용 소환 임박		lexrank	160.5054932
일만 에 침묵 갠 박 대통령 심경 토로 의혹이광호 유기준 일 만 침묵 갠 박근혜 대통령		textrank	103.6849899
박 대통령 새해 첫날 깜짝 기자 간담회 무특검 언론 불리한 여론 조성 불 표출 시각		lexrank	99.92407227
박 나 를 완전히 엮은 것 최순실 뇌물죄 박 대통령 최순실과 공모 여부 대해서 분명		textrank	94.6966095
박 대통령 뇌물죄 완전히 엮은 것 세월호 새해 첫날 청와대서 사실상 간담회 직무정지		textrank	105.7216034
박 삼성 합병 지시 혐의 에 완전히 엮은 박근혜 대통령 몇 십년 지인 박근혜 대통령		lexrank	90.22803497
특검 정유라 학점 특혜 소설 가 이인화 구조교 답안 대신 작성 부당 학점 의혹 특혜		textrank	196.7536621
박 대통령 도 사적 영역 있다간담회 성사 박근혜 대통령 대통령 모든 사람 자기 사적		lexrank	80.65332031

< 최종 데이터 >

### 3. Modeling 과정

---

#### 1. 데이터 전처리

##### 1-2. word tokenize & normalization

###### - tokenize

: 문장을 띄어쓰기 단위로 나눠 문장을 단어화

< Extract Summarization >

박 대통령은 먼저 "국민들께 미안한 생각으로 무거운 마음으로 지내고 있다"면서 "하루 빨리 안정을 찾아 나라가 발전의 탄력을 키워야 한다는 마음"이라며 국회 탄핵 가결 이후 직무정지된 상황에 대한 심경을 토로했다.



< tokenize >

[ 박, 대통령은, 먼저, 국민들께, 미안한, 생각으로, 무거운, 마음으로, 지내고, 있다면서, 하루, 빨리, 안정을, 찾아, 나라가, 발전의, 탄력을, 키워야, 한다는, 마음이라며, 국회, 탄핵, 가결, 이후, 직무정지된, 상황에, 대한, 심경을, 토로했다 ]

## 3. Modeling 과정

---

### 2. Modeling

#### 2-1. word & index dictionary

: look-up table에서 encoder input & decoder input에 해당하는 벡터를 불러오기 위해  
각 word에 고유한 index를 지정한 **word to index** 사전 생성  
또한, 반대되는 **index to word** 사전을 생성해 decoder의 output 벡터를 단어로 변환한 결과 도출

✓ 단어가 기준점(threshold)보다 빈도수가 작으면 word to index 사전에 포함하지 않음  
Extract 요약(text)에 존재하는 단어 순서대로 index 지정

✓ <UNK> : 기준점(threshold)보다 빈도수가 낮은 단어  
<PAD> : encoding size & decoding size보다 size가 작을 경우 채워주는 token  
<GO> : decoder input의 처음에 붙이는 신호  
<EOS> : target의 마지막에 붙이는 신호

[ 촌, 대통령, 국민, 께, 미안한, 생각, 무거운, 마음, 으로, 지내 ] → [2212, 97, 579, 523, 2213, 60, 766, 674, 2214, 1617]

## 3. Modeling 과정

---

### 2. Modeling

#### 2-2. Look-up table 초기값

: 빈도수 3이상인 단어에 대한 word2vec의 결과인 300차원 word vector로 look-up table 초기값 생성  
(word to index의 순서대로 word vector 나열)

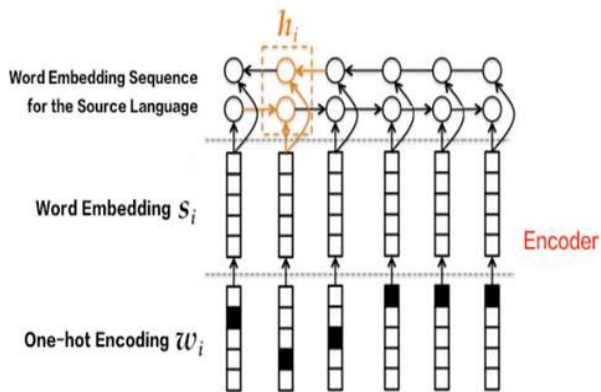
Look-up table: 각 행이 특정 word에 대한 word vector로 이루어진 array

$$\begin{pmatrix} \text{대통령(97)} : & [-0.04721257, -0.06758873, \dots] \\ \text{국민(579)} : & [0.11269119, -0.04476307, \dots] \\ & \circ \\ & \circ \\ & \circ \\ \text{朴(2212)} : & [-0.01550575, -0.10764064, \dots] \end{pmatrix}$$

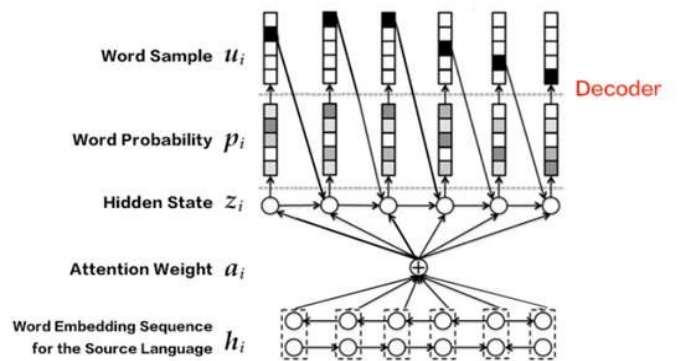
### 3. Modeling 과정

#### 2. Modeling

#### 2-3. Seq2Seq using BiLSTM with attention



< Encoder using BiLSTM >



< Decoder with Attention >

## 4. 결론 및 제언

## 4. 결론 및 제언

### 1. 결론

#### 1-1. E2A 모델 결과 (train set)

文 대통령 '세계 시민상' 수상

17.09.14 서울신문

애틀랜틱 카운슬 “민주주의 기여”...트뤼도 총리·음악가 랑랑 뽑혀

문재인 대통령이 국제협력·분쟁해결 분야의 세계적 연구기관인 애틀랜틱 카운슬이 주는 2017 [세계시민상](#)을 받았다.  
(... 중략)

문 대통령은 다음주 유엔총회 참석차 뉴욕을 방문할 때 애틀랜틱 카운슬이 주관하는 2017 세계시민상 [시상식](#)에 참석해 상을 받을 예정이다.

#### < 문재인 대통령 세계 시민상 수상 >

출처: [http://www.seoul.co.kr/news/newsView.php?id=20170915008014&wlog\\_tag3=naver#csidxa4fd19ce07ccda2903f4f464b0db72b](http://www.seoul.co.kr/news/newsView.php?id=20170915008014&wlog_tag3=naver#csidxa4fd19ce07ccda2903f4f464b0db72b)



## 4. 결론 및 제언

### 1. 결론

#### 1-2. E2A 모델 결과 (test set)

위안부 할머니 손잡고 이동하는 문재인 대통령 18.01.04 경향신문

문재인 대통령이 4일 오후 청와대 본관에서 휠체어에 앉은  
일본군 위안부 피해 할머니의 손을 잡고 오찬장으로 이동하고 있다.

< 위안부 할머니 손잡는 문 대통령 내외 >



출처: [http://news.khan.co.kr/kh\\_news/khan\\_art\\_view.html?artid=201801041540001&code=910203](http://news.khan.co.kr/kh_news/khan_art_view.html?artid=201801041540001&code=910203)

## 4. 결론 및 제언

### 1. 결론

#### 1-2. E2A 모델 결과 (test set)

'6월 항쟁 주역' 文대통령, 동시대 영화 '1987' 관람	18.01.07 경기일보
<p>6월항쟁 당시 인권 변호사로 민주화 운동을 위해 노력했던 문재인 대통령이 관련 시대적 배경을 그린 영화 '1987'을 7일 관람했다.</p> <p>(... 중략)</p> <p>문 대통령은 지난해 8월 5·18 광주항쟁을 그린 '택시운전사'를 관람했고, 10월에는 부산국제영화제에 참석해 워킹맘의 애환을 담은 '미씽, 사라진 여자'를 관람한 바 있다.</p>	

#### < 문 대통령 영화 택시 관람 >

출처: <http://www.kyeonggi.com/?mod=news&act=articleView&idxno=1430628>

## 4. 결론 및 제언

### 1. 결론

#### 1-3. 기존 model(기사 본문 데이터) vs E2A model(extract 요약 데이터)

: 동일한 train set으로 학습한 후, 동일한 test set에 대한 결과 비교

기존 model	E2A model	실제 기사 제목
문재인 대통령 집권 국무회 의 주재 공무원	문재인 대통령 첫 국무회 의	문재인 대통령 새해 첫 국무회 의 주재
왕세제 최측근 칼둔 특사 로 왕세제 해명	임종석 방문 국교 파트너십 강화 할 듯	칼둔 행정청장 내일 방한 문 대통령 임 실장 과 면담 예측
한미 군사훈련 연기 있는 文 대통령 트럼프와 함	文 대통령 트럼프와 통화 북핵 해법	남북 해빙무드 '성큼'... 힘받는 文 '운전대론'
文 대통령 내주 초 일정 없이	文 대통령 노동계 신년 인사회	문재인 대통령, 이달 중순 청와대서 중소기업인들 만나다

## 4. 결론 및 제언

---

### 1. 결론

1-3. 기존 model(기사 본문 데이터) vs E2A model(extract 요약 데이터)  
: 동일한 train set으로 학습한 후, 동일한 test set에 대한 결과 비교

	기존 model	E2A model
BLEU	0.5300	0.6209
ROUGE	0.2792	0.3574

## 4. 결론 및 제언

### 2. 제언

#### 2-1. 활용 방안(해시태그 자동완성)

: 제목이 없는 SNS 글을 대상으로 Abstract 요약을 생성해 해시태그를 자동완성



< 제천 화재현장 소방관  
격려 하는 문 대통령 >

# 제천

# 화재현장

# 소방관

#격려

# 文 대통령

출처: <https://www.facebook.com/TheBlueHouseKR/>

## 4. 결론 및 제언

---

### 2. 제언

#### 2-2. 보완점

- 1) 데이터 증가를 통한 model 성능 보완
- 2) 기존 Extract 요약 알고리즘이 아닌 독자적인 알고리즘 개발
- 3) Extract 요약 과정의 모델화를 통한 Extract 요약 & Abstract 요약 전체 모델의 단일화

**감사합니다.**