Hierarchical Attention Networks for Document Classification

2019. 07. 12

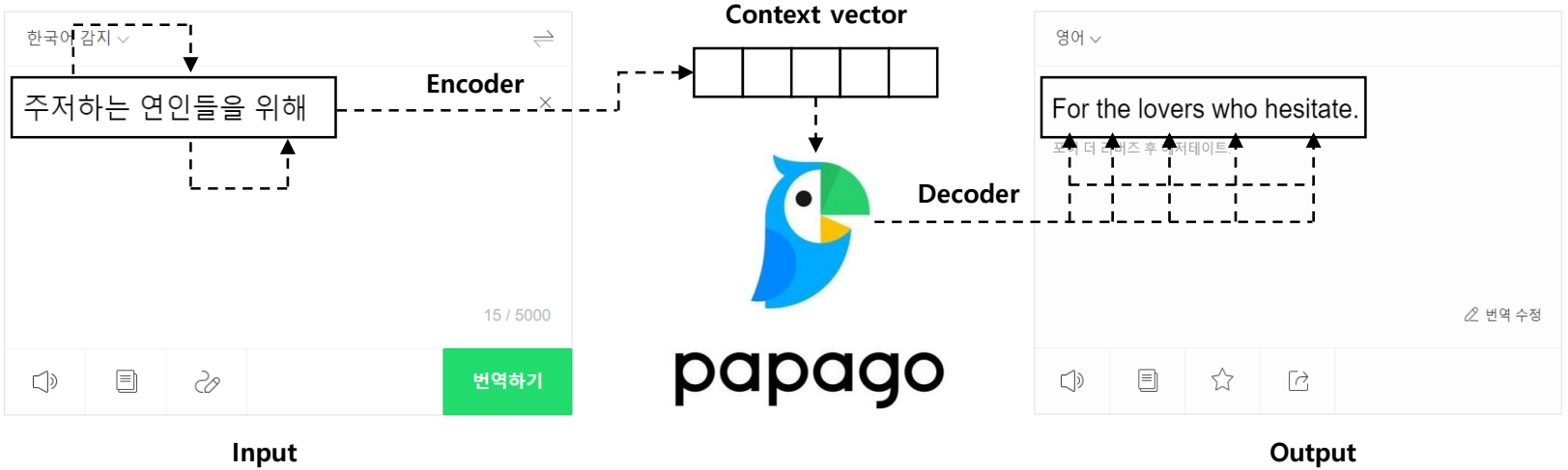
Data Mining & Quality Analytics Lab.

강 현 규

Contents

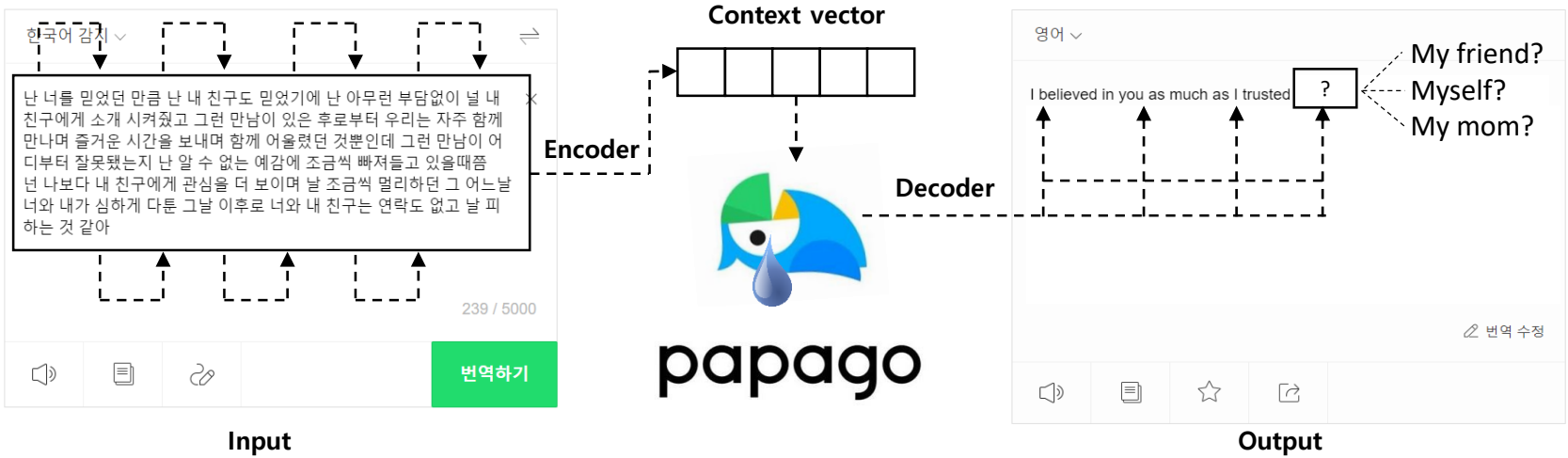
❖Attention Mechanism

- Introduction : Machine Translation



❖Attention Mechanism

- Introduction : Machine Translation



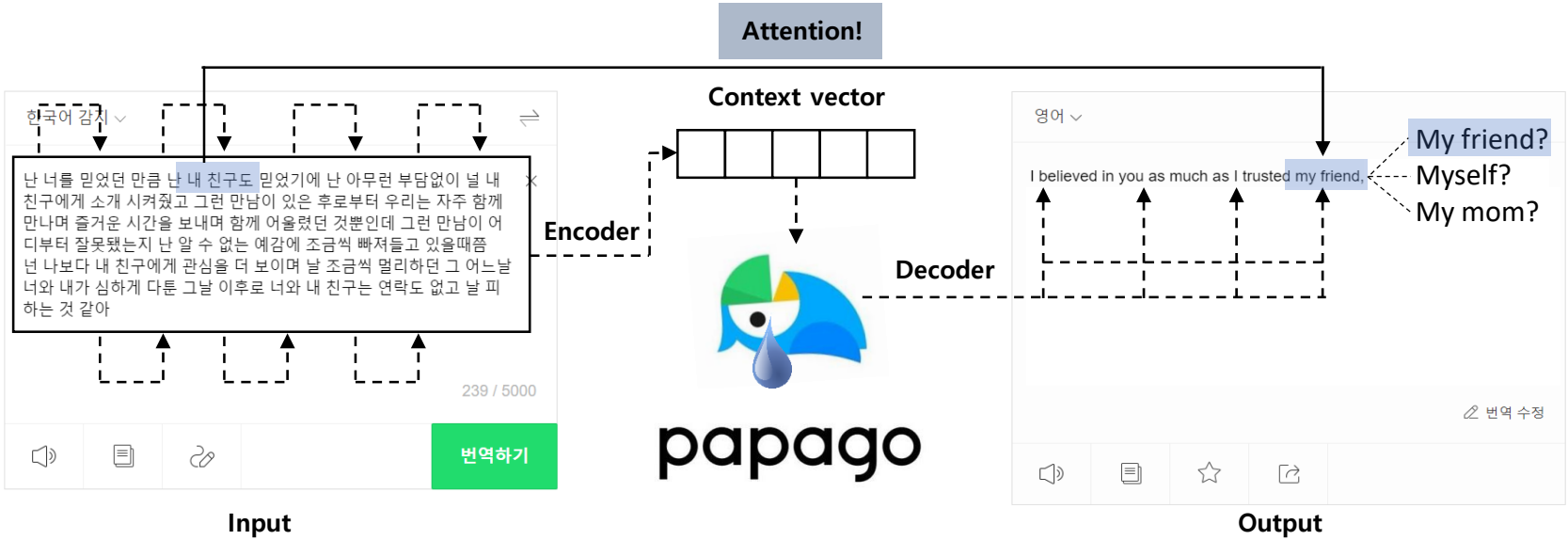
➢ **Long Input/output sequence**

→ Long term dependency

→ Vanishing gradient

❖Attention Mechanism

- Introduction : Machine Translation



➢ Motivation of Attention

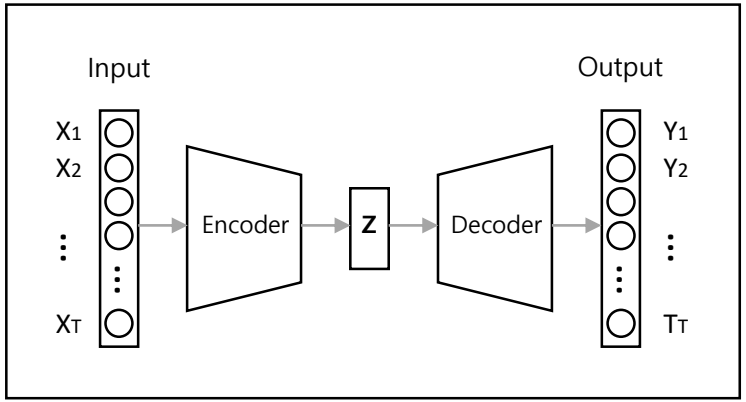
→ Seq2seq의 decoder에서 t시점의 단어를 예측할 때 decoder의 t-1 시점의 정보와 유사도가 높은 정보를 갖는 encoder의 단어를 주목하게 만든다

❖Attention Mechanism

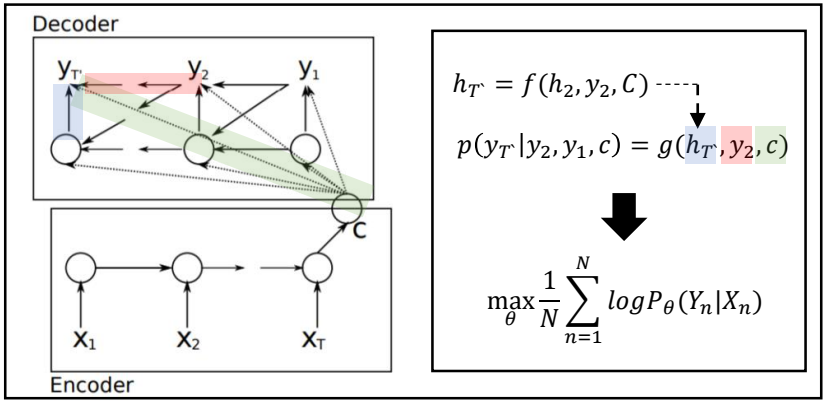
- RNN Encoder-Decoder(Cho et al, 2014)

➢ Summary

- AutoEncoder를 기계번역에 적용할 수 있도록 encoder/decoder에 DNN이 아닌 RNN layer을 사용
  ∵ 기계 번역 문제에서 pair sentence의 sequence는 가변
- LSTM의 구조를 단순화한 GRU layer을 제안



AutoEncoder



RNN – Encoder-Decoder

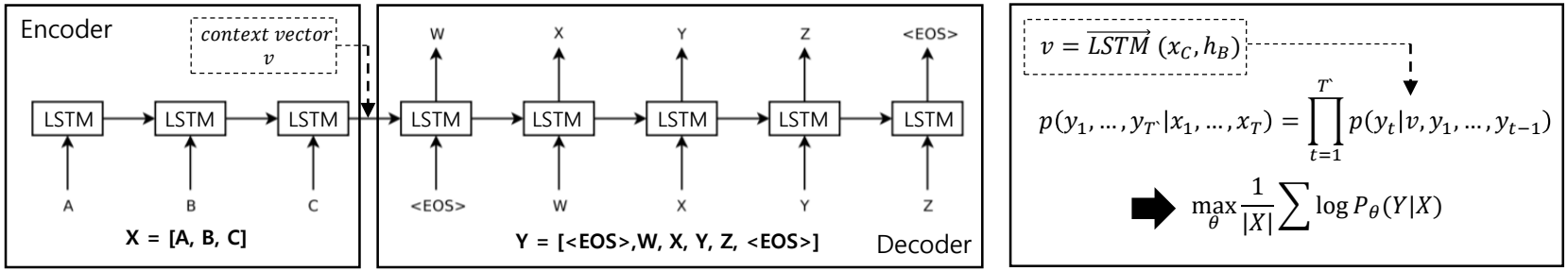
Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

❖ Attention Mechanism

- Seq2seq Model(Sutskever et al, 2014)

➢ Summary

- Seq2seq 용어가 공식적으로 등장
- Sequence에 <EOS> Token 사용
- LSTM layer
- Encoder의 문장 순서를 역순으로 학습(English-French 번역 모델)



Seq2seq Model

- RNN Encoder-Decoder 모델과 동일한 컨셉(Encoder의 마지막 RNN layer의 output = Context vector)
  → Long term dependency / Gradient vanishing

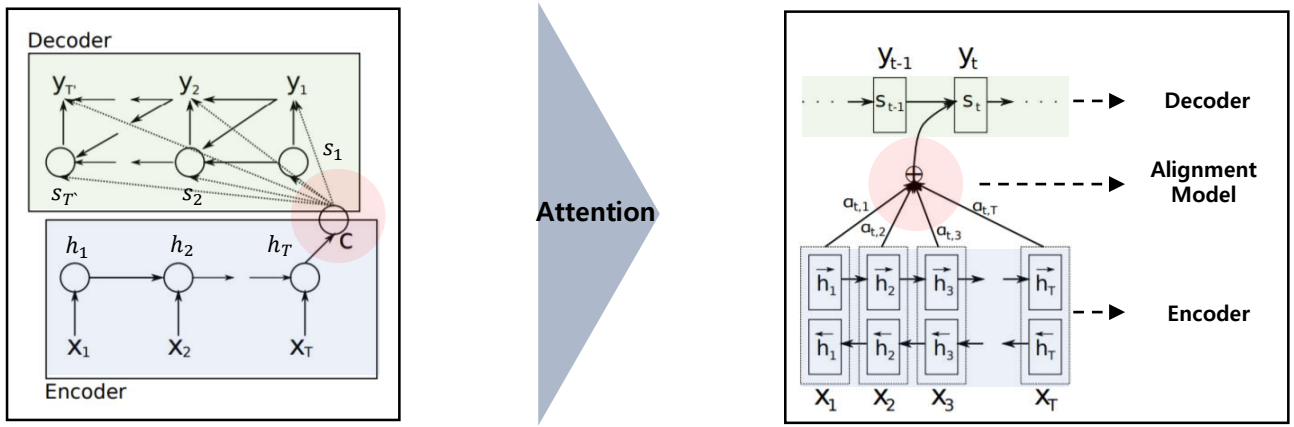
Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

❖Attention Mechanism

- Seq2seq Model with Attention(2014), Bahdanau attention

➢ Motivation of Attention

- Decoder가 단어를 예측할 때 encoder의 정보를 하나의 context vector으로 요약하는 것은 충분하지 않다
  → t 시점에서 사용하는 context vector을 c가 아닌 c_t를 사용하여 output을 예측하자



Seq2seq

$$p(y_t|y_{t-1},\dots,y_1,c) = g(s_t, y_{t-1}, c)$$

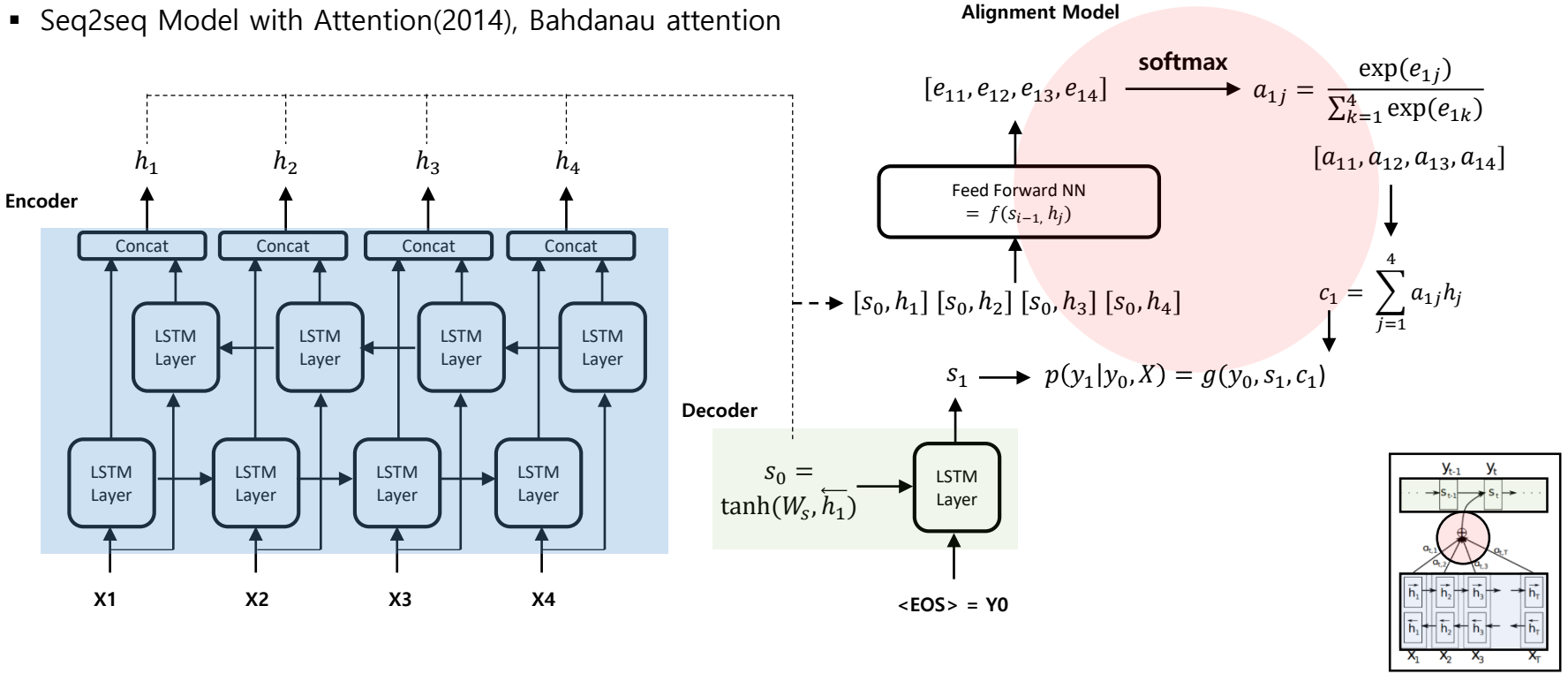
Seq2seq with Attention

$$p(y_t|y_{t-1},\dots,y_1,c) = g(s_t, y_{t-1}, c_t),\ c_t = \sum_{i=1}^{T} a_i h_i$$

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

❖Attention Mechanism

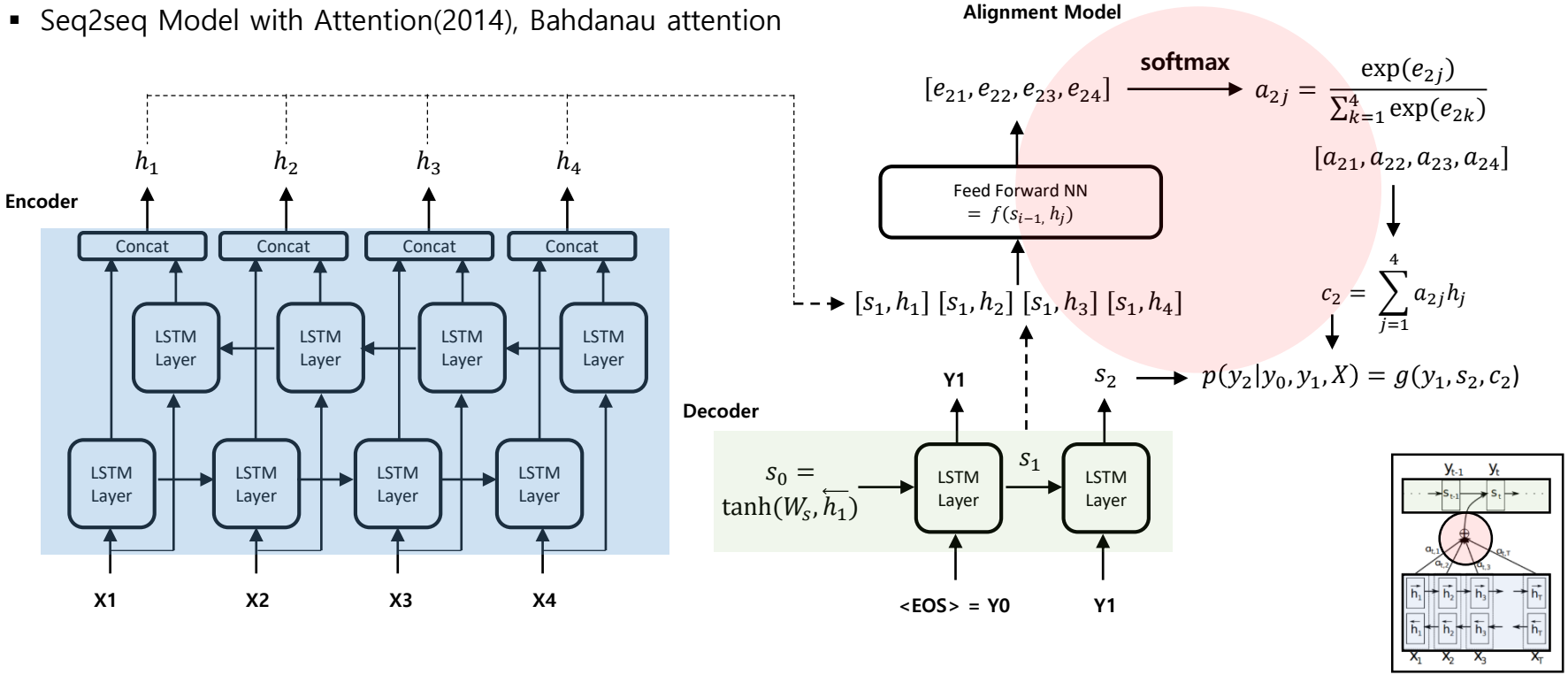
- Seq2seq Model with Attention(2014), Bahdanau attention



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

❖Attention Mechanism

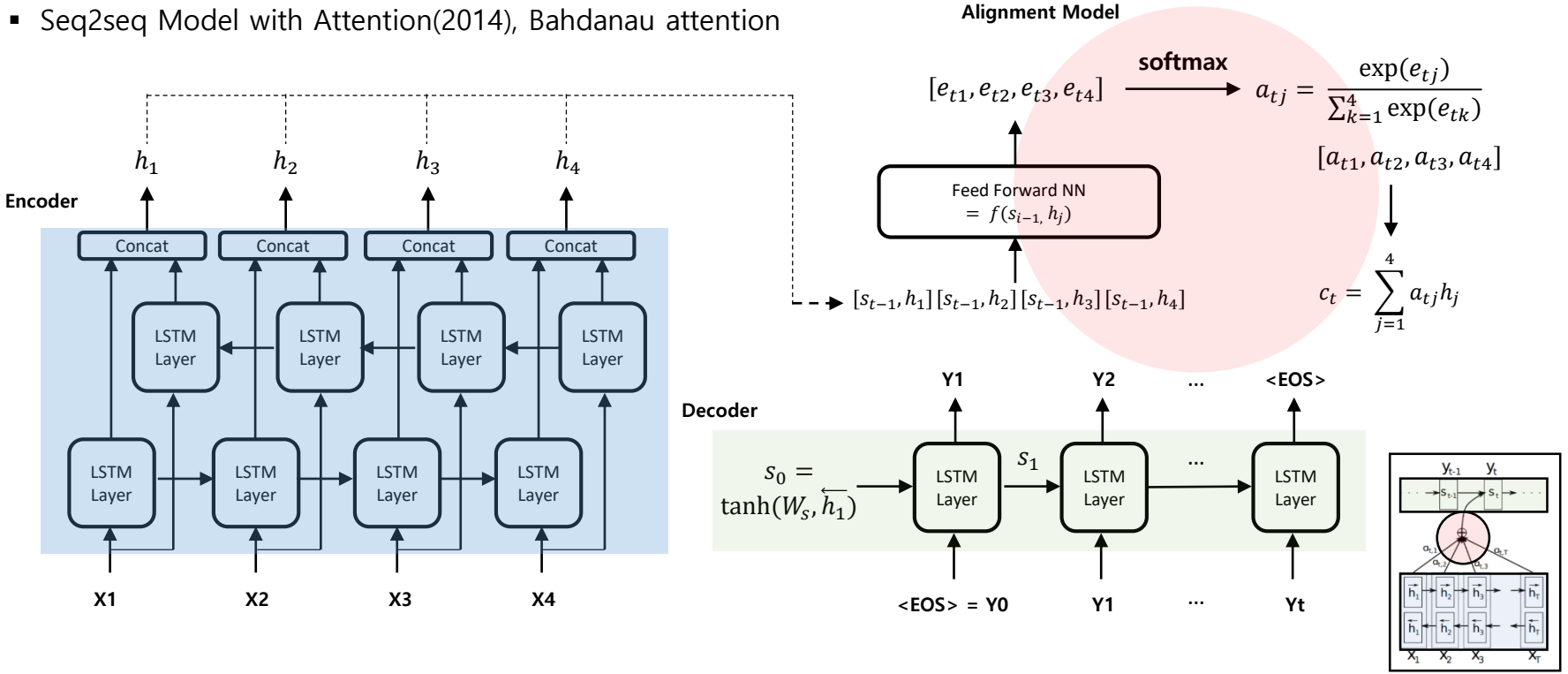
- Seq2seq Model with Attention(2014), Bahdanau attention



Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

❖ Attention Mechanism

- Seq2seq Model with Attention(2014), Bahdanau attention

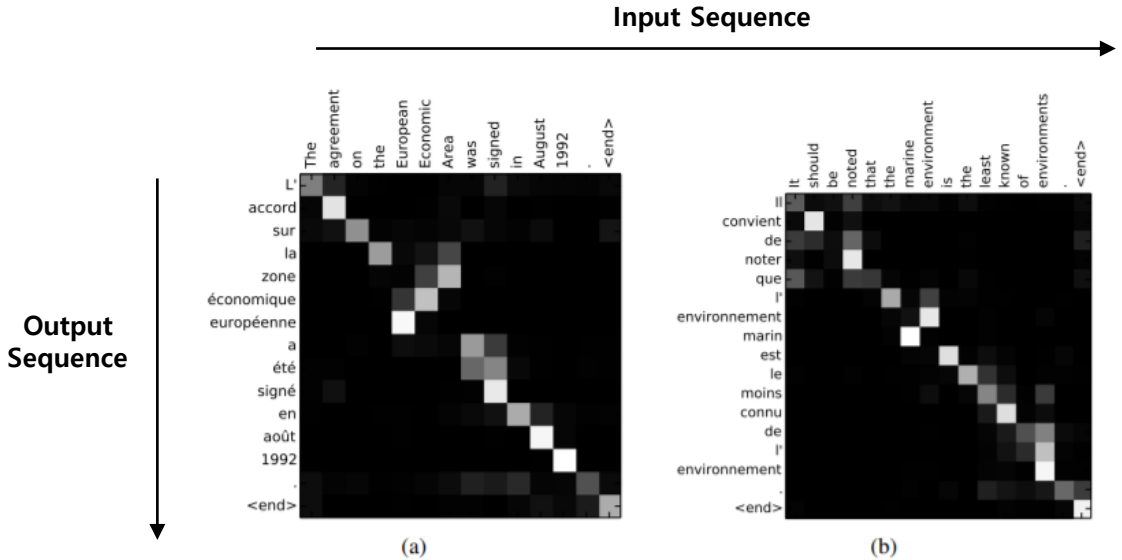


Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

❖ Attention Mechanism

- Seq2seq Model with Attention(2014), Bahdanau attention

➢ Alignment Matrix



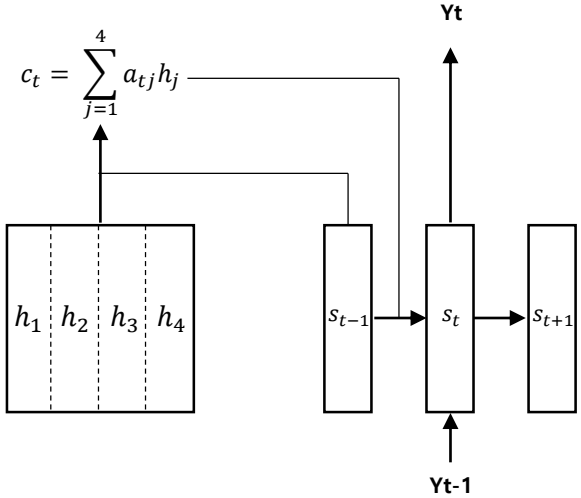
Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

❖Attention Mechanism

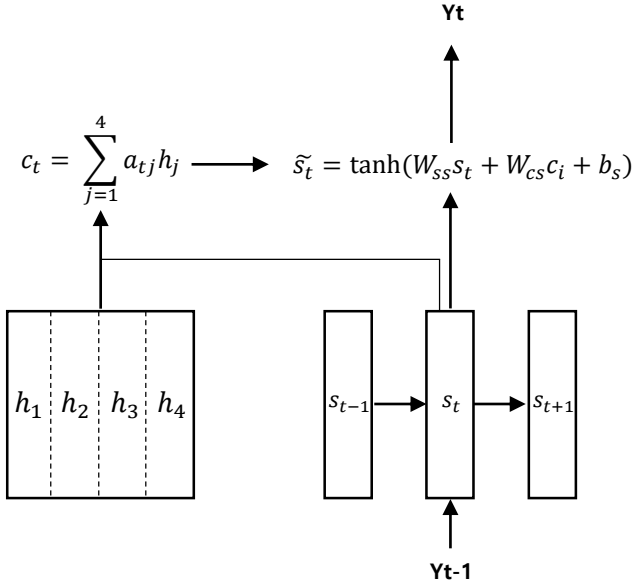
- Seq2seq Model with Attention(2015), Luong Attention

➢ Luong Attention

- Context vector 연산 시 s_{t-1}이 아닌 s_t 사용
- Attention score 연산 시 computation cost ↓



Bahdanau Attention



Luong Attention

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

❖ Hierarchical Attention Network

- Introduction : Document Classification

❖Hierarchical Attention Network

- Introduction : Document Classification

지금부터 2016헌나1 대통령 박근혜 탄핵 사건에 대한 선고를 시작하겠습니다. 저희 재판관들은 지난 90여 일 동안 이 사건을 공정하고 신속하게 해결하기 위하여 온 힘을 다하여 왔습니다.

(...)

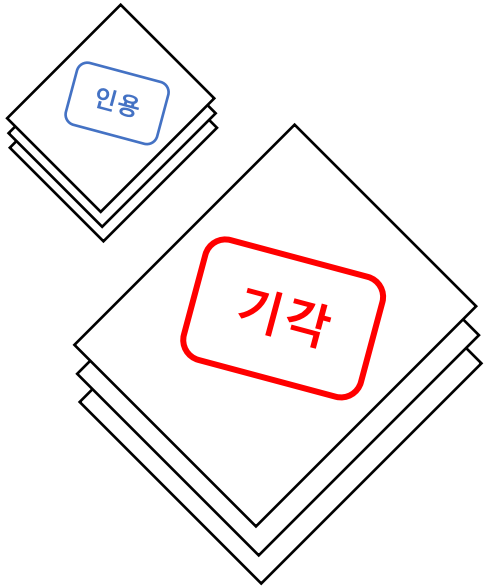
대통령비서실장 김기춘이 문화체육관광부 제1차관에게 지시하여 1급 공무원 여섯 명으로부터 사직서를 제출 받아 그 중 세 명의 사직서가 수리된 사실은 인정됩니다. 그러나 이 사건에 나타난 증거를 종합하더라도, **피청구인이 노 국장과 진 과장이 최서원의 사익 추구에 방해가 되었기 때문에 인사를 하였다고 인정하기에는 부족하고**, 유진룡이 면직된 이유나 **김기춘이 여섯 명의 1급 공무원으로부터 사직서를 제출 받도록 한 이유 역시 분명하지 아니합니다.**

(...)

피청구인은 국가가 국민의 생명과 신체의 안전 보호의무를 충실하게 이행할 수 있도록 권한을 행사하고 직책을 수행하여야 하는 의무를 부담합니다. 그러나 국민의 생명이 위협받는 재난상황이 발생하였다고 하여 **피청구인이 직접 구조 활동에 참여하여야 하는 등 구체적이고 특정한 행위의무까지 바로 발생한다고 보기는 어렵습니다.**

(...)

결국 피청구인의 위헌·위법행위는 국민의 신임을 배반한 것으로 헌법수호의 관점에서 용납될 수 없는 중대한 법 위배행위라고 보아야 합니다. 피청구인의 법 위배행위가 헌법질서에 미치는 부정적 영향과 파급효과가 중대하므로, 피청구인을 파면함으로써 얻는 헌법 수호의 이익이 압도적으로 크다고 할 것입니다.



❖Hierarchical Attention Network

- Introduction : Document Classification

지금부터 2016헌나1 대통령 박근혜 탄핵 사건에 대한 선고를 시작하겠습니다. 저희 재판관들은 지난 90여 일 동안 이 사건을 공정하고 신속하게 해결하기 위하여 온 힘을 다하여 왔습니다.

(...)

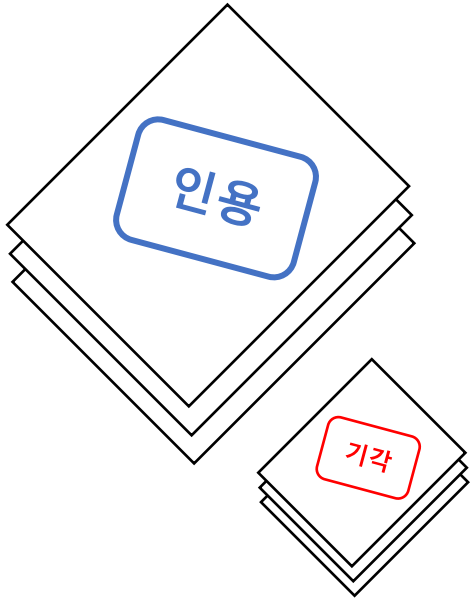
대통령비서실장 **김기춘이 문화체육관광부 제1차관에게 지시하여 1급 공무원 여섯 명으로부터 사직서를 제출 받아 그 중 세 명의 사직서가 수리된 사실은 인정됩니다.** 그러나 이 사건에 나타난 증거를 종합하더라도, 피청구인이 노 국장과 진 과장이 최서원의 사익 추구에 방해가 되었기 때문에 인사를 하였다고 인정하기에는 부족하고, 유진룡이 면직된 이유나 김기춘이 여섯 명의 1급 공무원으로부터 사직서를 제출 받도록 한 이유 역시 분명하지 아니합니다.

(...)

피청구인은 국가가 국민의 생명과 신체의 안전 보호의무를 충실하게 이행할 수 있도록 권한을 행사하고 직책을 수행하여야 하는 의무를 부담합니다. 그러나 국민의 생명이 위협받는 재난상황이 발생하였다고 하여 피청구인이 직접 구조 활동에 참여하여야 하는 등 구체적이고 특정한 행위의무까지 바로 발생한다고 보기는 어렵습니다.

(...)

결국 **피청구인의 위헌·위법행위는 국민의 신임을 배반한 것으로 헌법수호의 관점에서 용납될 수 없는 중대한 법 위배행위**라고 보아야 합니다. **피청구인의 법 위배행위가 헌법질서에 미치는 부정적 영향과 파급효과가 중대하므로, 피청구인을 파면**함으로써 얻는 헌법 수호의 이익이 압도적으로 크다고 할 것입니다.



❖Hierarchical Attention Network

- Introduction : Document Classification

지금부터 2016헌나1 대통령 박근혜 탄핵 사건에 대한 선고를 시작하겠습니다. 저희 재판관들은 지난 90여 일 동안 이 사건을 공정하고 신속하게 해결하기 위하여 온 힘을 다하여 왔습니다.

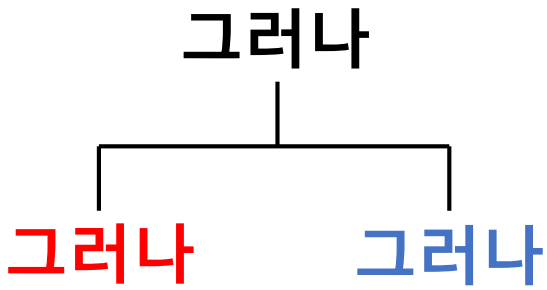
(...)

대통령비서실장 김기춘이 문화체육관광부 제1차관에게 지시하여 1급 공무원 여섯 명으로부터 사직서를 제출 받아 그 중 세 명의 사직서가 수리된 사실은 인정됩니다. **그러나 이 사건에 나타난 증거를 종합하더라도, 피청구인이 노 국장과 진 과장이 최서원의 사익 추구에 방해가 되었기 때문에 인사를 하였다고 인정하기에는 부족하고, 유진룡이 면직된 이유나 김기춘이 여섯 명의 1급 공무원으로부터 사직서를 제출 받도록 한 이유 역시 분명하지 아니합니다.**

(...)

대통령은 헌법과 법률에 따라 권한을 행사하여야 함은 물론, 공무 수행은 투명하게 공개하여 국민의 평가를 받아야 합니다. **그러나 피청구인은 최서원의 국정개입사실을 철저히 숨겼고, 그에 관한 의혹이 제기될 때마다 이를 부인하며 오히려 의혹 제기를 비난하였습니다.** 이로 인해 국회 등 헌법기관에 의한 견제나 언론에 의한 감시 장치가 제대로 작동될 수 없었습니다. 또한, 피청구인은 미르와 케이스포츠 설립, 플레이그라운드와 더블루케이 및 케이디코퍼레이션 지원 등과 같은 최서원의 사익 추구에 관여하고 지원하였습니다.
피청구인의 헌법과 법률 위배행위는 재임기간 전반에 걸쳐 지속적으로 이루어졌고, 국회와 언론의 지적에도 불구하고 오히려

(...)



❖Hierarchical Attention Network

- Introduction : Document Classification

지금부터 2016헌나1 대통령 박근혜 탄핵 사건에 대한 선고를 시작하겠습니다. 저희 재판관들은 지난 90여 일 동안 이 사건을 공정하고 신속하게 해결하기 위하여 온 힘을 다하여 왔습니다.

(...)

대통령비서실장 김기춘이 문화체육관광부 제1차관에게 지시하여 1급 공무원 ○○명으로부터 사직서를 제출 받아 그 중 ○명의 사직서가 수리된 사실은 인정됩니다. 그러나 이 사건에 나타난 증거를 종합하더라도, 피청구인이 노 국장과 진 과장이 최서원의 사익 추구에 방해가 되었기 때문에 인사를 하였다고 인정하기에는 부족하고, 유진룡이 면직된 이유나 김기춘이 여섯 명의 1급 공무원으로부터 사직서를 제출 받도록 한 이유 역시 분명하지 아니합니다.

(...)

피청구인은 국가가 국민의 생명과 신체의 안전 보호의무를 충실하게 이행할 수 있도록 권한을 행사하고 직책을 수행하여야 하는 의무를 부담합니다. 그러나 국민의 생명이 위협받는 재난상황이 발생하였다고 하여 피청구인이 직접 구조 활동에 참여하여야 하는 등 구체적이고 특정한 행위의무까지 바로 발생한다고 보기는 어렵습니다.

(...)

[illegible] 피청구인의 위헌·위법행위는 국민의 신임을 배반한 것으로 헌법수호의 관점에서 용납될 수 없는 중대한 법 위배행위라고 보아야 합니다. 피청구인의 법 위배행위가 헌법질서에 미치는 부정적 영향과 파급효과가 중대하므로, 피청구인을 파면함으로써 얻는 헌법 수호의 이익이 압도적으로 크다고 할 것입니다.

"The intuition underlying our model is that ***not all parts of a document are equally relevant for answering a query*** and that determining the relevant sections involves modeling **the *interactions of the words, not just their presence in isolation***."

- 문서 분류에 있어 문서를 구성하는 문장/단어가 갖는 중요도가 다름
- 같은 단어여도 문맥에 따라 의미하는 바가 다름

❖ Hierarchical Attention Network

- Introduction : Document Classification

지금부터 2016헌나1 대통령 박근혜 탄핵 사건에 대한 선고를 시작하겠습니다. 저희 재판관들은 지난 90여 일 동안 이 사건을 공정하고 신속하게 해결하기 위하여 온 힘을 다하여 왔습니다.

(...)

대통령비서실장 김기춘이 문화체육관광부 제1차관에게 지시하여 1급 공무원 여섯 명으로부터 사직서를 제출받아 그 중 세 명의 사직서가 수리된 사실은 인정됩니다. 그러나 이 사건에 나타난 증거를 종합하더라도, 피청구인이 노 국장과 진 과장이 최서원의 사익 추구에 방해가 되었기 때문에 인사를 하였다고 인정하기에는 부족하고, 유진룡이 면직된 이유나 김기춘이 여섯 명의 1급 공무원으로부터 사직서를 제출 받도록 한 이유 역시 분명하지 아니합니다.

(...)

피청구인은 국가가 국민의 생명과 신체의 안전 보호의무를 충실하게 이행할 수 있도록 권한을 행사하고 직책을 수행하여야 하는 의무를 부담합니다. 그러나 국민의 생명이 위협받는 재난상황이 발생하였다고 하여 피청구인이 [illegible] 여야 하는 등 구체적이고 특정한 행위의무 [illegible] 바로 발생한다고 보기는 어렵습니다. [illegible] 피청구인의 위헌·위법행위는 국민의 신임을 배반한 것으로 헌법수호의 관점에서 용납될 수 없는 중대한 법 위배행위라고 보아야 합니다. 피청구인의 법 위배행위가 헌법질서에 미치는 부정적 영향과 파급효과가 중대하므로, 피청구인을 파면함으로써 얻는 헌법 수호의 이익이 압도적으로 크다고 할 것입니다.

"The intuition underlying our model is that ***not all parts of a document are equally relevant for answering a query*** and that determining the relevant sections involves modeling **the interaction of the words, not just their presence in isolation**."

- 문서 분류에 있어 문서를 구성하는 문장/단어가 갖는 중요도가 다름
- 같은 단어여도 문맥에 따라 의미하는 바가 다름

How to make better document representation for classification?

❖Hierarchical Attention Network

- Introduction : Document Classification

➢ **Hierarchical Attention Network for Document Classification**

- Yang, Zichao, et al
- Carnegie Mellon University / Microsoft
- 2016 NAACL Conference
  (북미컴퓨터언어학학회)

[PDF] **Hierarchical attention** networks for document classification
Z Yang, D Yang, C Dyer, X He, A Smola... - Proceedings of the 2016 ..., 2016 - aclweb.org
We propose a **hierarchical attention** network for document classification. Our model has two distinctive characteristics:(i) it has a **hierarchical** structure that mirrors the **hierarchical** structure of documents;(ii) it has two levels of **attention** mechanisms applied at the wordand ...
☆ 99 997회 인용 관련 학술자료 전체 10개의 버전 »

Hierarchical Attention Networks for Document Classification

Zichao Yang[1], Diyi Yang[1], Chris Dyer[1], Xiaodong He[2], Alex Smola[1], Eduard Hovy[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond
{zichaoy, diyiy, cdyer, hovy}@cs.cmu.edu
xiaohe@microsoft.com alex@smola.org

Abstract

We propose a hierarchical attention network for document classification. Our model has two distinctive characteristics: (i) it has a hierarchical structure that mirrors the hierarchical structure of documents; (ii) it has two levels of attention mechanisms applied at the word- and sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation. Experiments conducted on six large scale text classification tasks demonstrate that the proposed architecture outperform previous methods by a substantial margin. Visualization of the attention layers illustrates that the model selects qualitatively informative words and sentences.

pork belly = delicious . || scallops? || I don't even like scallops, and these were a-m-a-z-i-n-g . || fun and tasty cocktails. || next time I in Phoenix, I will go back here. || Highly recommend.

Figure 1: A simple example review from Yelp 2013 that consists of five sentences, delimited by period, question mark. The first and third sentence delivers stronger meaning and inside, the word *delicious*, *a-m-a-z-i-n-g* contributes the most in defining sentiment of the two sentences.

Although neural-network–based approaches to text classification have been quite effective (Kim, 2014; Zhang et al., 2015; Johnson and Zhang, 2014; Tang et al., 2015), in this paper we test the hypothesis that better representations can be obtained by

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

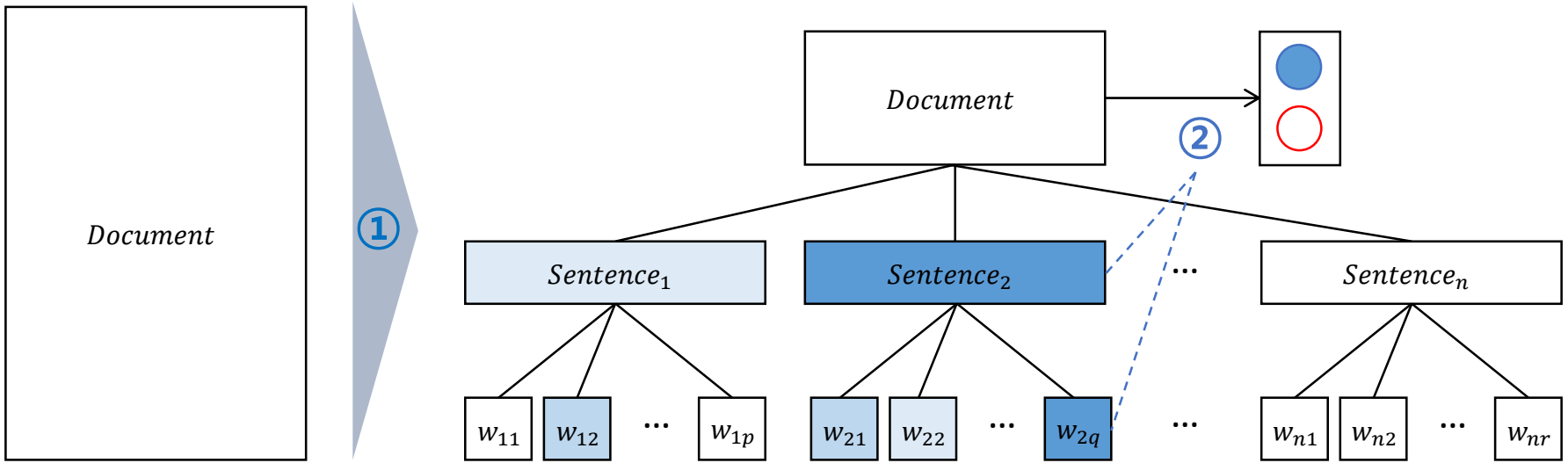
❖ Hierarchical Attention Network

- Introduction : Document Classification

➢ **Key Idea for Document Representation**

① 문서를 표현함에 있어 문서 – 문장 –단어의 계층 구조를 활용하자

② Hierarchical Attention을 통해 문서의 중요 정보를 갖는 문장과 단어의 정보를 학습에 더 반영할 수 있도록 하자



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ **Key Idea for Document Representation**

① 문서를 표현함에 있어 문서 – 문장 –단어의 계층 구조를 활용하자

② Hierarchical Attention을 통해 문서의 중요 정보를 갖는 문장과 단어의 정보를 학습에 더 반영할 수 있도록 하자

➢ **Hierarchical Structure**

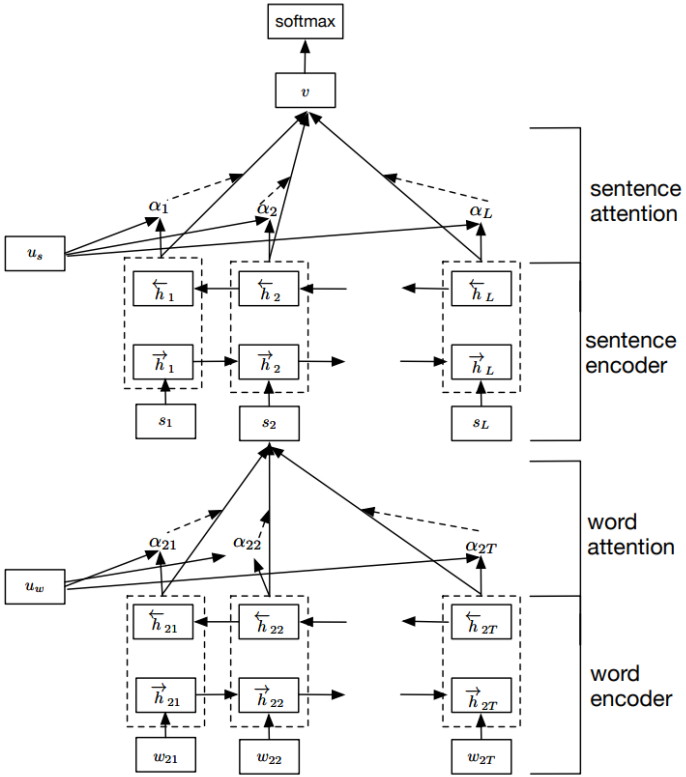
① Word Sequence Encoder

② Word-level Attention

③ Sentence Encoder

④ Sentence-level Attention

⑤ Classifier

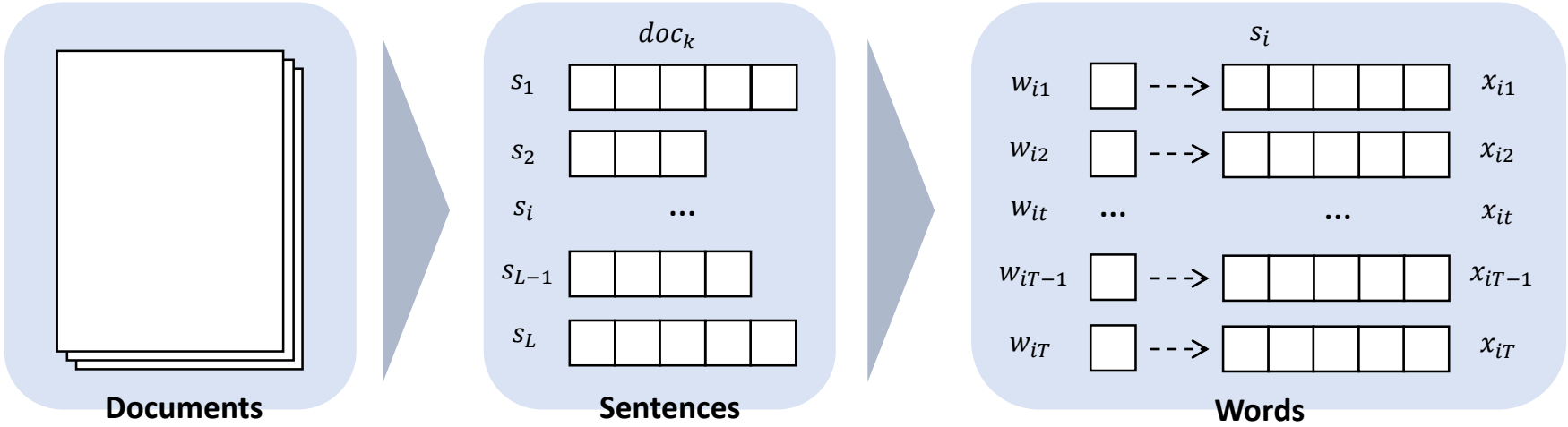


Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Terminology



- ✓ $L =$ # of sentences in document k
- ✓ $T =$ # of words in sentence i
- ✓ $w_{it} = t^{th}$ word in sentence i
- ✓ $x_{it} =$ Embeding vector of w_{it}

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Hierarchical Structure

① Word Sequence Encoder

- Word Embedding Layer

$$x_{it} = W_e w_{it}, t \in [1, T]$$

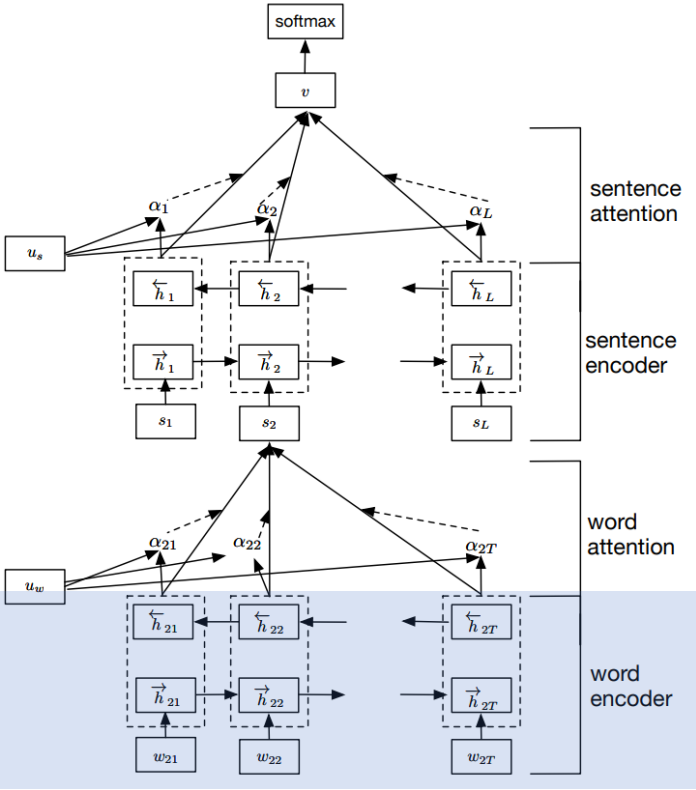
(w_{it}: Distributed representation of word)

- Word Encoder(Bidirectional GRU Layer)

$$\vec{h}_{it} = \overrightarrow{GRU}(x_{it}), t \in [1, T]$$

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [1, T]$$

concat → $h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$ → output



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Hierarchical Structure

② Word Attention

$$h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$$

↓ Pass one-layer MLP with tanh activation

$$u_{it} = \tanh(W_w h_{it} + h_w) \dashrightarrow \text{Hidden representation of word}$$

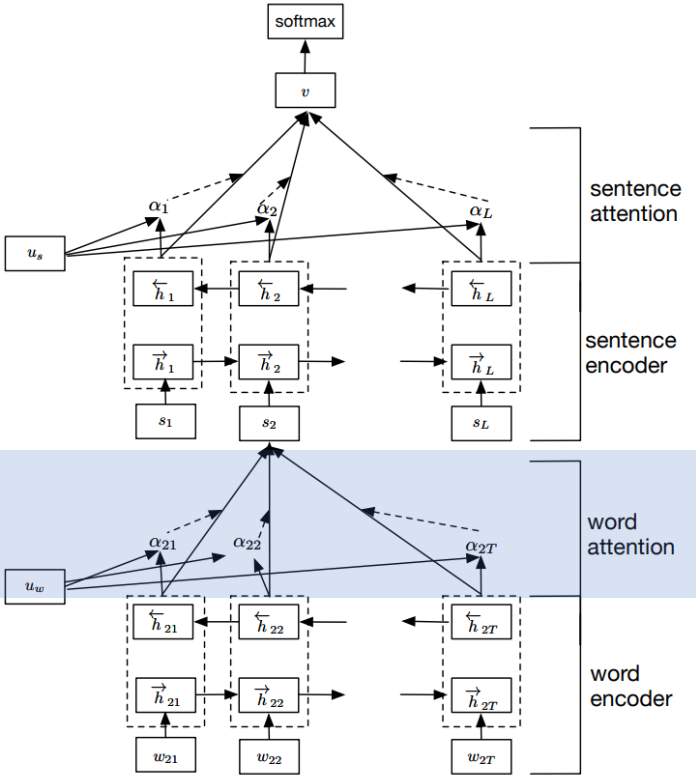
↓ Dot product

$$u_{it}^T u_w \dashrightarrow \text{Similarity between word and word context vector}$$

↓ Softmax

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \dashrightarrow \text{Normalized importance weight}$$

= Attention score of word



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Hierarchical Structure

③ Sentence Encoder

$$s_i = \sum_t \alpha_{it} h_{it}$$ ⟶ Weight sum of word vectors = Sentence representation

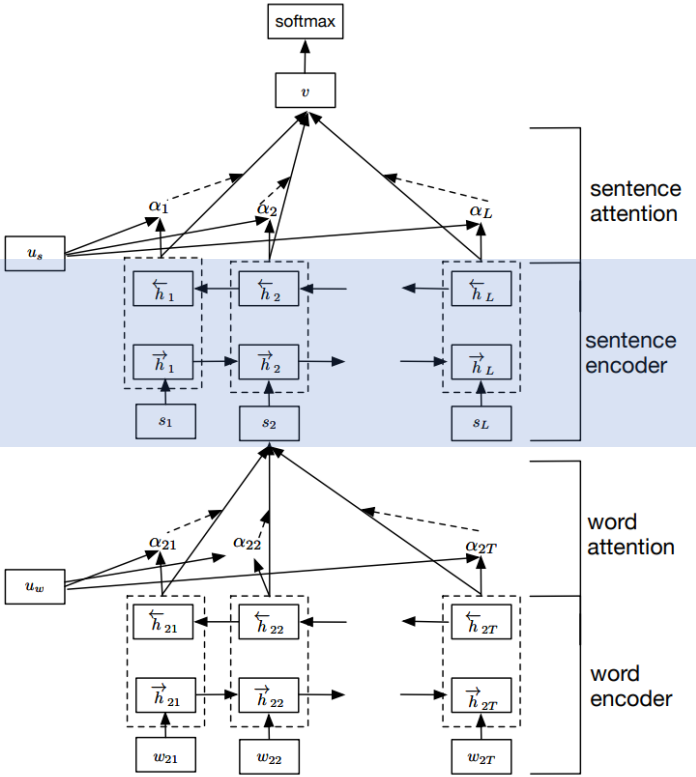
- Sentence Encoder(Bidirectional GRU Layer)

$$\vec{h}_i = \overrightarrow{GRU}(s_i), t \in [1, L]$$

$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i), t \in [1, L]$$

Concat ⟶ $h_i = [\vec{h}_i, \overleftarrow{h}_i]$

Output



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Hierarchical Structure

④ Sentence Attention

$$h_i = [\vec{h}_i, \ \overleftarrow{h}_i]$$

↓ Pass one-layer MLP with tanh activation

$$u_i = \tanh(W_s h_i + h_s)$$ ---→ Hidden representation of sentence

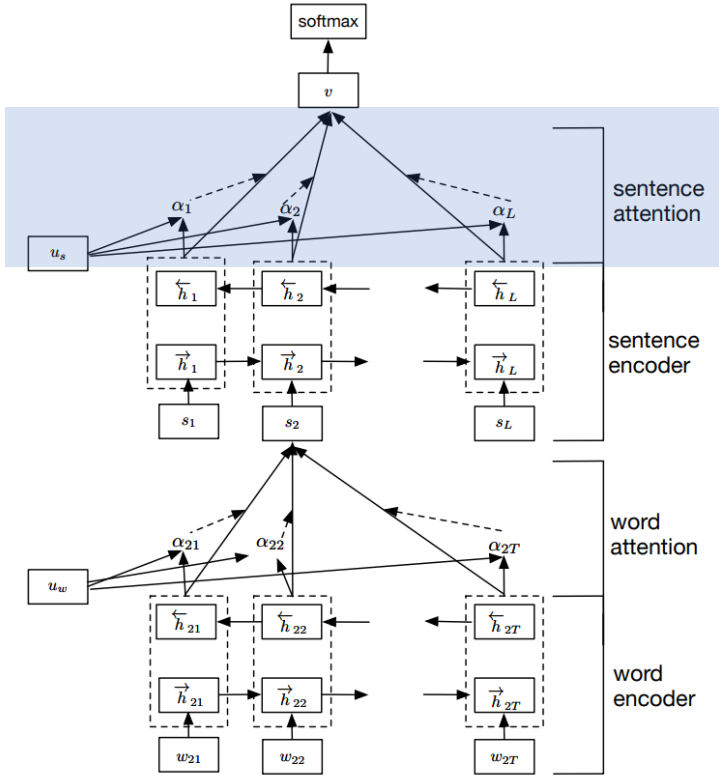
↓ Dot product

$$u_i^T u_s$$ ----------→ Similarity between sentence and sentence context vector

↓ Softmax

$$\alpha_i = \frac{\exp(u_i^T u_w)}{\sum_t \exp(u_i^T u_w)}$$ -----→ Normalized importance weight

= Attention score of sentence



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Model Architecture

➢ Hierarchical Structure

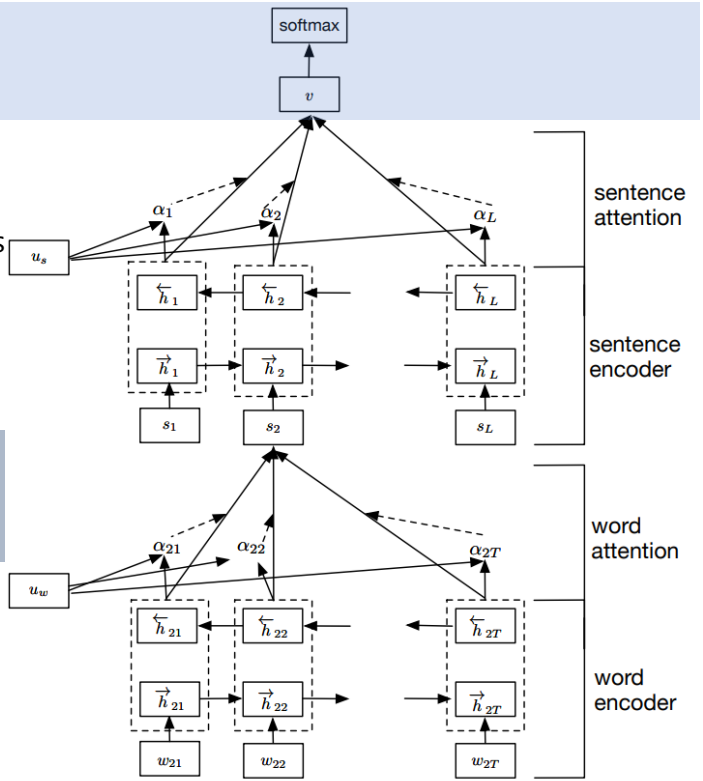
⑤ Classifier

$v = \sum_i \alpha_i h_i$ ----------→ Weight sum of sentence vectors = Document represenetation

$p = softmax(W_c v + b_c)$ ---→ Probability of document class

$(W_c = Class\ weight)$

$Training\ Loss = -\sum_d \log(p_{dj})$ -→ Minimize negative log-likelihood of the correct labels



Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Experiment

➢ Experiment Details

① Dataset

1) Yelp : 서비스 업장 리뷰, 5점 척도
2) IMDB : 영화 리뷰, 10점 척도
3) Amazon review : 상품 리뷰, 5점 척도
4) Yahoo Answer : 질의 응답, 10개 카테고리

------→ Multi-class classification

Data set	classes	documents	average #s	max #s	average #w	max #w	vocabulary
Yelp 2013	5	335,018	8.9	151	151.6	1184	211,245
Yelp 2014	5	1,125,457	9.2	151	156.9	1199	476,191
Yelp 2015	5	1,569,264	9.0	151	151.9	1199	612,636
IMDB review	10	348,415	14.0	148	325.6	2802	115,831
Yahoo Answer	10	1,450,000	6.4	515	108.4	4002	1,554,607
Amazon review	5	3,650,000	4.9	99	91.9	596	1,919,336

② Process

1) Dataset split : Training : Validation : Test = 8 : 1 : 1
2) Word embedding W_e : Initialized by using Word2vec
3) Context vector u_w, u_s : Initialized randomly
4) Evaluation : Classification Accuracy

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Experiment

➢ Experiment Details

③ Result

*" the neural network-based methods that **do not explore hierarchical document structure**, such as LSTM, CNN-word, CNN char **have little advantage over traditional methods** for large scale text classification"*

- Non hierarchical NN-based model 은 전통적인 Count-based model과 분류 성능에 큰 차이를 보이지 않음
- 문서 분류 문제에서는 결국 특정 단어의 등장 빈도가 분류를 판단

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

Table 2: Document Classification, in percentage

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Experiment

➢ Experiment Details

③ Result

*"Exploring **the hierarchical structure only**, as in HN-AVE, HAN-MAX can significantly improve over LSTM, CNN-word, and CNN-Char*

(..)

*Compared to HN-AVE, **the HN-ATT model gives superior performance** across the board"*

- 문서의 구조적인 특성을 반영했을 때 이전에 비해 분류 성능이 향상
- Attention mechanism이 분류 성능 향상에 큰 기여

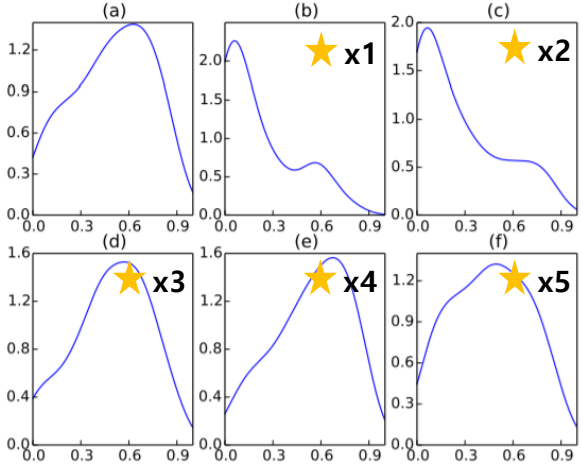
	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

Table 2: Document Classification, in percentage

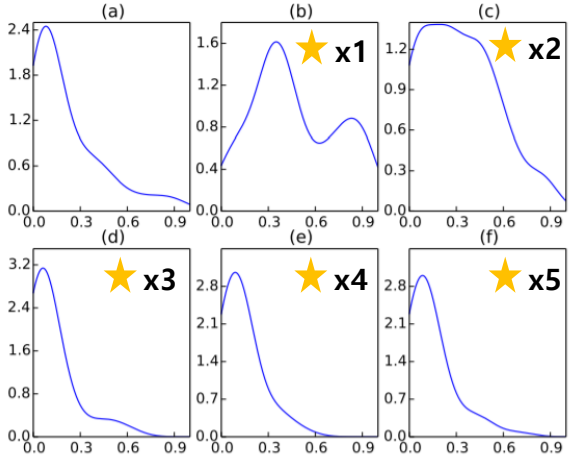
Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Etc : Context Dependent Attention Weights



[Attention Score Distribution of 'good']



[Attention Score Distribution of 'bad']

같은 단어라도 다른 문맥에서 사용되었을 때의 Attention score의 분포가 다름

★ **x1** : Not good ★ **x5** : Very good -----➢ $Dist(\alpha_{good}) < Dist(\alpha_{good})$

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Hierarchical Attention Network

- Etc : Visualization of Attention

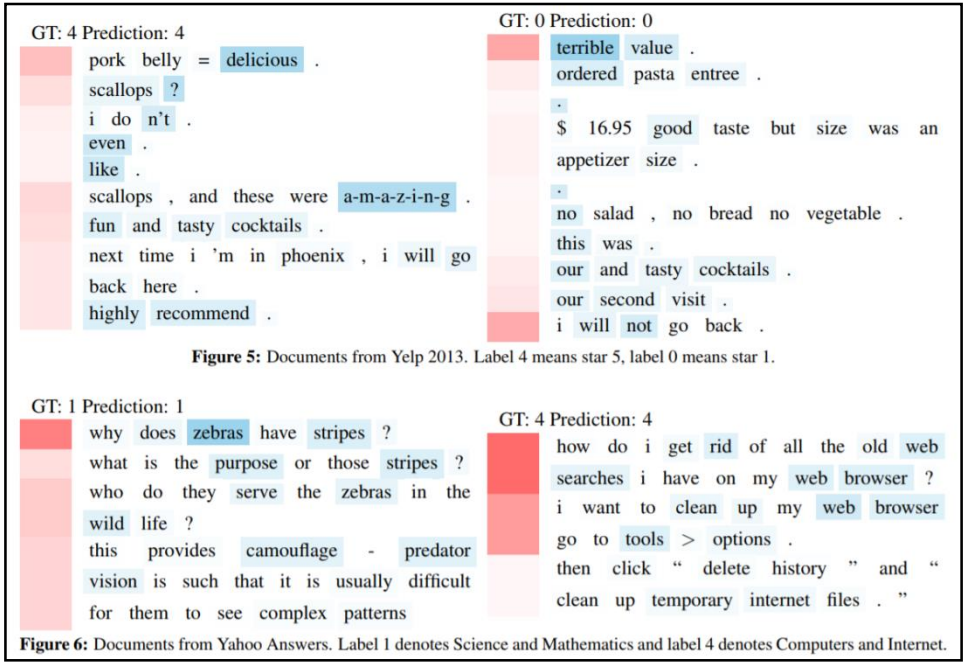


Figure 5: Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.

Figure 6: Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.

: Sentence attention score

: Word attention score

- 문서의 Class를 분류하는데 있어 중요한 정보를 담고 있는 문장과 단어를 attention score으로 확인

Yang, Zichao, et al. "Hierarchical attention networks for document classification." *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

❖ Conclusion

➢ Contribution

① 문서의 구조적인 특성을 반영하여 분류 성능을 향상

② 문맥을 반영하여 단어와 문장의 의미를 학습

➢ Limitation : RNN + RNN

① Computational cost

② Long term dependency

➢ Personal Review

① 간단한 아이디어를 모델로 구체화시켜 분류 성능 향상

② 간결하고 읽기 쉬운 논문

➢ But…

❖ Conclusion

➢ State-of-the-Art Models in NLP(Including Classification Task)

- Self – Attention : Break away from RNN CNN Architecture
- Pre-Trained & Fine Tuning

❖ Reference

➢ Paper

- Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
- Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).
- Yang, Zichao, et al. "Hierarchical attention networks for document classification." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016.

➢ Blog

- http://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
- https://lovit.github.io/machine%20learning/2019/03/17/attention_in_nlp/
- https://hcnoh.github.io/2019-01-01-luong-attention

감사합니다