MASS와 복사 메커니즘을 이용한 한국어 문서 요약

정영준[O1], 이창기[1], 고우영[2], 윤한준[2]

[1]강원대학교 컴퓨터과학과, [2]ETRI 부설 연구소
{kongjun, leeck}@kangwon.ac.kr, {gwy876, hjyoon}@nsr.re.kr

Korean Text Summarization using MASS with Copying Mechanism

Young-Jun Jung[O1], Chang-Ki Lee[1], Woo-Young Go[2], Han-Jun Yoon[2]
[1]Department of Computer Science, Kangwon National University, [2]The Affiliated Institute of ETRI

요 약

문서 요약(text summarization)은 주어진 문서로부터 중요하고 핵심적인 정보를 포함하는 요약문을 만들어 내는 작업으로, 기계 번역 작업에서 주로 사용되는 Sequence-to-Sequence 모델을 사용한 end-to-end 방식의 생성(abstractive) 요약 모델 연구가 활발히 진행되고 있다. 최근에는 BERT와 MASS 같은 대용량 단일 언어 데이터 기반 사전학습(pre-training) 모델을 이용하여 미세조정(fine-tuning)하는 전이 학습(transfer learning) 방법이 자연어 처리 분야에서 주로 연구되고 있다. 본 논문에서는 MASS 모델에 복사 메커니즘(copying mechanism) 방법을 적용하고, 한국어 언어 생성(language generation)을 위한 사전학습을 수행한 후, 이를 한국어 문서 요약에 적용하였다. 실험 결과, MASS 모델에 복사 메커니즘 방법을 적용한 한국어 문서 요약 모델이 기존 모델들보다 높은 성능을 보였다.

주제어: 문서 요약, 사전학습, MASS, 복사 메커니즘

1. 서론

문서 요약(text summarization)은 주어진 문서로부터 중요하고 핵심적인 정보를 포함하는 요약문을 만들어내는 작업으로, 문서 요약 작업에 사용되는 접근법은 크게 두 가지로 구분할 수 있다. 첫 번째는 추출(extractive) 요약으로, 입력된 문서에 있는 문장 중에서 핵심적인 문장을 선택하고 추출하여 요약문으로 사용하는 방법이다. 추출 요약은 완전한 문장으로 구성된 문서에서 문장을 선택하기 때문에 불완전한 문장이 만들어지지 않는 장점이 있지만, 문서 내에 중요한 문장이 없을 경우 문제가 될 수 있으며, 요약문의 응집도나 가독성이 떨어질 수 있다. 두 번째는 생성(abstractive) 요약으로, 문서 내용을 이해하여 새로운 요약문을 생성하는 언어 생성(language generation) 작업이다. 생성 요약은 불완전하거나 부자연스러운 문장이 만들어질 가능성이 있어 추출 요약보다 어려운 방법이지만, 더 간결하고 함축적인 문장을 생성할 수 있는 장점이 있다.

생성 요약은 같은 언어 생성 작업인 기계 번역 작업에서 주로 사용되는 Sequence-to-Sequence 모델을 사용한 end-to-end 방식의 모델이 주로 연구되고 있다[1,2]. 일반적으로 요약문에는 요약에 사용한 원본 문서에 있는 단어들이 나타나는 경우가 많은데, Sequence-to-Sequence 모델을 사용하여 요약문을 생성할 경우 고유명사와 같은 단어들의 생성 확률이 낮은 문제점이 있다. 복사 메커니즘(copying mechanism)은 이러한 문제를 해결하고자 제안된 방법으로, 디코딩 과정에서 단어를 생성할 확률과 입력 문서에서 단어를 복사할 확률에 가중치를 부여하게 되고, 가중치에 따라 고유명사와 같은 생성 확률이 낮은 단어도 입력 문서에서 찾아 복사할 수 있게 된다[3].

언어 생성 작업은 주로 지도 학습(supervised learning) 방법을 사용하는데, 일반적으로 지도 학습 방법은 데이터가 부족할 경우에는 낮은 성능을 보인다. 따라서 최근에는 학습 데이터 부족 문제를 해결하기 위해 대량의 단일 언어 데이터를 이용하여 사전학습(pre-training)하는 방법을 사용하는 BERT(Bidirectional Encoder Representations from Transformers)[4]와 MASS(Masked Sequence to Sequence pre-training)[5] 같은 언어 모델(language model)이 자연어 처리 분야에서 많은 관심을 끌고 있다.

본 논문에서는 다양한 언어 생성 작업에서 좋은 성능을 보이고 있는 MASS 모델에 복사 메커니즘 방법을 적용하여 한국어 문서 요약 모델의 성능이 개선될 수 있음을 보인다.

2. 관련 연구

Sequence-to-Sequence 모델은 인코더와 디코더로 구성된 모델로, 인코더는 입력 열을 인코딩하고 디코더는 인코딩된 정보를 받아 출력 열을 만들어낸다. [1]에서 제안된 모델은 어텐션 메커니즘(attention mechanism) 기반 인코더-디코더(encoder-decoder) 모델이다. 어텐션 메커니즘은 출력 단어를 예측하기 위해 집중해서 봐야 할 입력 문장의 단어에 대한 어텐션 가중치를 결정한다. [2]에서는 순환(recurrence)과 합성곱(convolution)을 없앤 단순한 어텐션 메커니즘 기반 인코더-디코더 구조의 트랜스포머(Transformer) 모델을 제안하였다. 트랜스포머는 멀티헤드 셀프어텐션(multi-head self-attention)을 사용하며, 언어 생성 작업에서 좋은 성능을 보이고 있다.

[3]은 디코딩 과정에서 단어를 생성할 확률과 입력 문

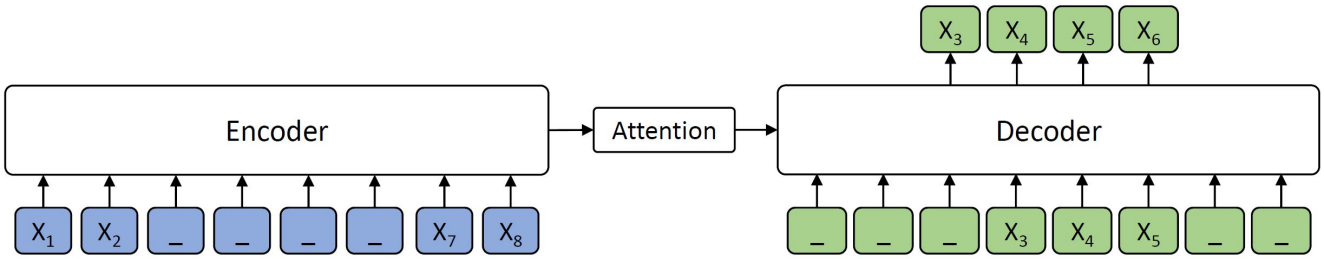

그림 1. MASS의 인코더-디코더 모델 구조[5]

서에서 단어를 복사할 확률에 가중치를 부여하는 복사 메커니즘과 이전 단어들을 생성하는 데에 사용한 어텐션 가중치를 고려하여 단어를 생성하는 커버리지 메커니즘(coverage mechanism)을 Sequence-to-Sequence 모델에 적용하여 문서 요약에서 좋은 성능을 보였다. [6]에서는 생성 요약에서 좋은 성능을 보여주고 있는 트랜스포머 모델에 복사 메커니즘과 추론 단계에서 페널티를 추가한 모델을 제안하고 이를 한국어 문서 생성 요약에 적용하여 좋은 성능을 보여주었다.

BERT[4]는 양방향 트랜스포머 인코더로 구성된 모델로, 문장 내 임의의 단어를 마스킹하고 예측하는 MLM(Masked Language Model)과 다음 문장 예측(Next Sentence Prediction)을 기반으로 모델을 사전학습한다. 사전학습된 BERT 모델은 다른 자연어 처리 작업에 미세조정(fine-tuning)하는 방법으로 적용되며, 이는 다양한 자연어 처리 작업에서 높은 성능을 보이고 있다. MASS[5]는 언어 생성 작업을 위해 인코더와 디코더를 공동으로 사전학습하는 모델로, 어텐션 메커니즘도 같이 사전학습되기 때문에 기계 번역과 문서 요약 같은 언어 생성 작업에서 좋은 성능을 보이고 있다[7].

3. MASS

MASS는 언어 생성 작업을 위해 인코더와 디코더를 함께 사전학습하는 모델로, 입력 문장 x가 주어지면 위치 u에서 위치 v까지 토큰이 마스크 되며, $0 < u < v < m$이다. 여기서 m은 문장 x의 토큰 개수이고, 위치 u에서 v까지 마스크 되는 토큰 수는 $k = v - u + 1$ 로 나타낼 수 있다. 마스크 된 토큰은 마스크 기호 [M]으로 대체되며, 마스크 된 문장의 길이는 바뀌지 않고 입력 문장과 동일한 길이를 가진다.

사전학습은 마스크 된 문장 x를 인코더의 입력으로 사용하며 디코더에서 위치 u에서 v까지 마스크 된 토큰을 예측하는 Sequence-to-Sequence 모델을 학습한다. 그림 1은 MASS 모델 구조의 예를 보여준다. 인코더에 $x_3x_4x_5x_6$ 토큰이 마스크 된 8개의 토큰을 가지는 문장이 입력되고, 디코더 입력으로는 위치 4-6의 토큰 $x_3x_4x_5$가 주어진다. 모델은 인코더의 입력 문장에서 마스크 된 토큰 $x_3x_4x_5x_6$만 예측하고, 디코더에서 위치 4-6을 제외한 다른 위치에 대한 입력으로는 특수 마스크 기호 [M]을 사용한다(위치 1-3, 7-8).

MASS는 언어 생성 작업을 위해 인코더와 디코더를 공동으로 사전학습하도록 설계되었다. Sequence-to-Sequence 모델을 통해 마스크 된 토큰만 예측하게 함으로써, 인코더가 마스크 되지 않은 토큰의 의미를 이해하도록 하고, 디코더는 인코더로부터 유용한 정보를 추출하도록 한다. 디코더에서는 연속적인 토큰을 예측하여 기존의 MLM 보다 언어 생성 작업에 더 적합한 언어 모델링을 할 수 있다. 또한, 마스크 되지 않은 디코더의 입력 토큰을 마스크 함으로써, 디코더는 이전 토큰의 정보를 활용하는 대신 인코더에서 유용한 정보를 더 추출하도록 한다.

기존의 사전학습 모델과 MASS의 차이를 마스크 된 토큰 길이를 나타내는 하이퍼파라미터(hyperparameter) k 값에 따라 다음과 같이 설명할 수 있다. $k = 1$인 경우, 입력 문장의 하나의 토큰만이 마스크 되며, 이는 BERT에서 사용되는 MLM과 같다. $k = m$인 경우, 인코더의 모든 토큰이 마스크 되고, 디코더는 이전 토큰으로 다음 토큰을 예측하는 일반적인 언어 모델링이 된다.

4. 복사 메커니즘

복사 메커니즘은 디코딩 과정에서 단어를 생성할 확률과 입력 문장에서 단어를 복사할 확률에 가중치를 부여하는 방법으로, 가중치에 따라 고유명사와 같은 생성 확률이 낮은 단어도 입력 문서에서 찾아 복사할 수 있게 된다. 본 논문에서는 언어 생성 작업에서 좋은 성능을 보이고 있는 인코더-디코더 모델인 트랜스포머 모델에 복사 메커니즘을 적용하였다. 그림 2는 복사 메커니즘을 적용한 트랜스포머 모델의 구조를 나타낸다.

복사 메커니즘의 어텐션 분포는 출력 단어를 예측하기 위해 집중해서 봐야 할 입력 문장의 단어에 대한 확률 분포를 나타낸다. 트랜스포머 모델에서 사용하는 스케일 내적 어텐션(scaled dot-product attention)을 통해 복사 메커니즘에서 사용하는 어텐션 분포 a^t를 계산한다. 단어를 생성할 확률과 복사할 확률에 대한 가중치를 부여하는 생성 확률 $p_{\text{gen}} \in [0,1]$는 식 (1,2)와 같이 계산된다.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

$$p_{\text{gen}} = \sigma\left(W_g\left[s_t, \text{Attention}\left(W_c^Q s_t, W_c^K h_i, W_c^V h_i\right)\right]\right) \quad (2)$$

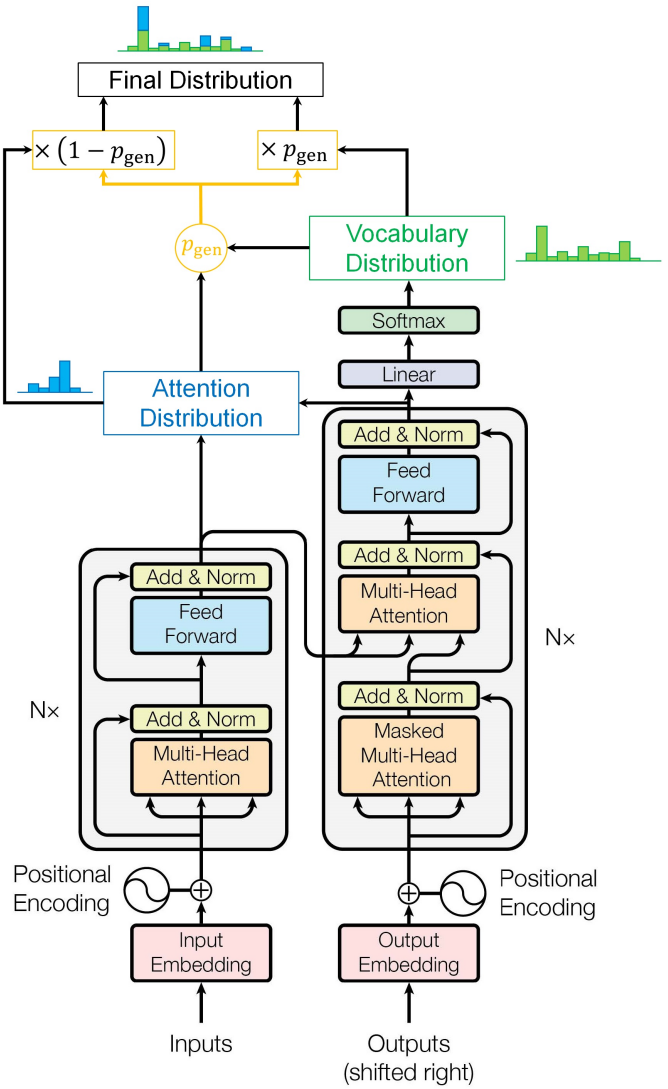

그림 2. 복사 메커니즘을 적용한 트랜스포머 모델 구조

d_k 는 쿼리(query)와 키(key)의 차원 수이고, W_c^Q, W_c^K, W_c^V, W_g는 학습 가중치, s_t는 디코더의 출력, h_i는 인코더의 히든 스테이트(hidden state), σ는 시그모이드(sigmoid) 함수를 나타낸다. 생성 확률 p_{gen}는 P_{vocab}의 어휘에서 단어를 생성하거나 a^t를 반영하여 입력 문장에서 단어를 복사하는 것 중에서 선택하도록 가중치를 부여하는 데 사용된다. 최종 단어 분포 $P(w)$는 p_{gen}와 P_{vocab}, a^t에 따라 식(3)과 같이 결정된다.

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t \qquad (3)$$

최종적으로 $P(w)$에서 가장 높은 확률값을 가지는 단어가 디코더의 출력 단어로 결정된다.

5. 실험

문서 요약 실험은 한국어에 대한 단일 언어 데이터를 사용해 MASS 모델을 사전학습하고, 한국어 문서 요약 데이터로 미세조정을 수행하여 진행하였다. 학습에 사용한 트랜스포머 모델의 하이퍼파라미터는 다음과 같다. 인코더와 디코더의 레이어 수는 6, 임베딩과 히든 레이어 차원 수는 768, 헤드 수는 12, 피드포워드(feed-forward) 차원 수는 3072이다. 실험에는 fairseq를 기반으로 구현된 MASS를 사용하였다[1]. 사전학습에서는 임의의 시작 위치 u를 가지는 연속적인 토큰을 [M]으로 교체하며, 마스크 길이 k를 문장의 총 토큰 수의 50%로 정하고 마스킹한다. BERT에서의 마스크 방법과 같이, 그중 80%는 마스크, 10%는 임의의 토큰으로 변경하고, 나머지 10%는 변경하지 않고 그대로 사용한다.

사전학습은 네이버 뉴스 크롤링 데이터 약 1,000만 문장을 한국어 학습 데이터로 사용하였다. 문서 요약 학습 데이터는 국립국어원 문서 요약 말뭉치[2]를 사용하였으며, 총 4,389개의 문서 데이터를 학습데이터 3,500개, 개발 데이터 449개, 평가 데이터 440개로 나누어 실험에 사용하였다. 모든 데이터는 형태소 분석기를 사용하여 형태소 단위로 분리한 뒤 BPE[8]를 적용하여 사용하였다. 표 1은 전처리 된 데이터에 대한 예제를 보여준다.

문서 요약 모델의 성능 평가 지표는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)를 사용하였다. ROUGE는 언어 생성 모델의 성능을 평가하기 위해 자주 이용되는 성능 평가 지표로, 정답 요약문(reference summary)과 시스템 요약문(candidate summary)간의 재현율(recall)을 바탕으로 모델의 성능을 평가한다. ROUGE-N은 정답 요약문과 시스템 요약문 간 중복되는 N-gram을 비교하는 방법으로, ROUGE-1은 유니그램(unigram), ROUGE-2는 바이그램(bigram)을 기준으로 성능을 측정하고, ROUGE-L은 최장 공통 부분 수열(LCS)을 이용해 성능을 측정한다.

표 2는 MASS와 복사 메커니즘을 이용한 한국어 문서 요약 실험에 대한 성능 결과를 나타내고, 표 3은 실험에 사용한 요약 모델들을 통해 생성한 요약문 예제이다. Transformer와 Transformer+Copy 모델은 사전학습을 진행하지 않고 문서 요약 데이터로 학습한 결과이고, Transformer+Copy 모델은 복사 메커니즘을 적용한 트랜스포머 모델의 성능이다. MASS, MASS+Copy, MASS+Copy(Both) 모델은 모두 MASS 사전학습을 진행하고 문서 요약 데이터로 미세조정한 모델의 성능이다. MASS+Copy와 MASS+Copy(Both) 모델의 차이점은 MASS+Copy 모델은 사전학습 과정에서 복사 메커니즘을 적용하지 않고 미세조정 과정에서만 복사 메커니즘을 적용한 모델이고, MASS+Copy(Both) 모델은 사전학습과 미세조정 과정에서 모두 복사 메커니즘이 적용된 모델이다.

실험 결과, MASS 모델에 복사 메커니즘을 적용한 모델들의 성능이 기존 모델들보다 우수한 성능을 보였다. 사전학습을 하지 않은 Transformer와 Transformer+Copy 모델을 비교하였을 때, 복사 메커니즘을 적용한 Transformer+Copy 모델이 높은 성능을 보였다. 이를 통해 복사 메커니즘이 문서 요약 작업에 적용되었을 때 요

[1] https://github.com/microsoft/MASS
[2] https://corpus.korean.go.kr

약 성능이 개선됨을 알 수 있다. 사전학습을 한 후에 미세조정을 한 MASS, MASS+Copy, MASS+Copy(Both) 모델을 비교하였을 때, MASS 모델보다 MASS+Copy 모델이 더 좋은 성능을 보였다. MASS+Copy 모델은 사전학습 과정에서 복사 메커니즘을 적용하지 않고 미세조정 과정에서만 적용하였지만, 복사 메커니즘을 사전학습하지 않고 미세조정 과정에서만 적용하여도 문서 요약 모델에 유의미한 영향을 주어 성능이 향상된 것으로 보인다. MASS+Copy(Both) 모델은 MASS+Copy 모델보다 더 좋은 성능을 보였다. 이는 사전학습 과정에서 복사 메커니즘도 같이 학습하였기 때문에 성능이 향상된 것으로 보인다. MASS 모델과 MASS+Copy, MASS+Copy(Both) 모델의 비교를 통해 사전학습을 사용하는 MASS 모델에 복사 메커니즘 방법을 적용하였을 때도 문서 요약 모델의 성능이 개선됨을 알 수 있다.

표 1. 한국어 문서 요약 데이터 예제

원본 데이터	여야가 6일 임시국회 종료 이틀을 남기고 주요 쟁점법안 처리 문제를 일괄 타결했다.
전처리 된 데이터	여야/NNG 가/JKS 6/SN 일/NNBC 임시/NNG 국회/NNG 종료/NNG 이틀/NNG 을/JKO 남기/VV 고/EC 주요/NNG 쟁점/NNG 법안/NNG 처리/NNG 문제/NNG 를/JKO 일괄/NNG 타결/NNG 했/XSV 다/EF ./SF

표 2. 한국어 문서 요약 결과

모델	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	40.26	14.81	62.49
Transformer+Copy	49.84	20.23	69.10
MASS	66.68	52.73	92.63
MASS+Copy	67.12	53.30	93.33
MASS+Copy(Both)	**67.45**	**53.81**	**94.06**

6. 결론

본 논문에서는 다양한 언어 생성 작업에서 좋은 성능을 보이고 있는 MASS 모델에 복사 메커니즘 방법을 적용하였다.

실험 결과, 대용량 단일 언어 데이터로 사전학습된 MASS 모델에 복사 메커니즘 방법을 적용한 모델이 기존의 문서 요약 모델보다 좋은 성능을 보였다. 또한, 복사 메커니즘을 사전학습하지 않고 미세조정 과정에서만 적용하여도 문서 요약 모델의 성능이 개선됨을 보였다.

향후 연구로는 문서 요약에 주로 사용되는 커버리지 메커니즘(coverage mechanism) 등의 기법들을 추가로 적용하여 모델을 개선할 예정이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원과 NSR의 지원을 받아 수행된 연구임 (No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발)

참고문헌

[1] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR*, 2015.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," *NIPS*, 2017.

[3] Abigail See, Peter J. Liu, Christopher D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," *arXiv preprint arXiv:1704.04368*, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, "MASS: Masked Sequence to Sequence Pre-training for Language Generation," *ICML*, 2019.

[6] 전동현, 강인호, "복사-메커니즘과 추론 단계의 페널티를 이용한 Copy-Transformer 기반 문서 생성 요약," *제31회 한글 및 한국어 정보처리 학술대회 논문집*, 2019.

[7] 정영준, 황현선, 이창기, "MASS를 이용한 한국어 문서 요약," *2019년 한국소프트웨어종합학술대회 논문집*, 2019.

[8] Rico Sennrich, Barry Haddow, Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units," *ACL*, 2016.

표 3. 한국어 문서 요약 예제

추출 주제 문장	마음을 다스리고 소송 당사자들과의 소통 능력을 높이기 위해 판사들도 '자기관리'를 한다. 판사 16명과 법원 직원 40명은 점심시간을 이용해 법원이 마련한 '사상체질과 스트레스' 강좌에 참여했다. 14일 오후 법원 6층에 마련된 요가실에서는 판사 4명이 스스로를 돌아보는 명상을 하고 있었다.
정답 요약 문장	의심이 직업병이 되고, 스트레스가 커진 판사들이 마음을 다스리고 소송 당사자들과의 소통 능력을 높이기 위해 '자기관리'를 하고 있다. 서울동부지법은 지난해에도 판사와 직원들을 대상으로 스트레스 해소 교육을 실시했는데 15일 낮 12시 신관 4층 대강당에서 열린 '사상체질과 스트레스' 강좌에 판사 16명과 법원 직원 40명이 참석했다. 서울북부지법 판사들은 지난달 2일부터 매주 월요일 오후 5시에 진행되는 '마음챙김 명상 프로그램' 참가하고 있고, 14일 오후 법원 6층에 마련된 요가실에서는 판사 4명이 명상을 하고 있었다.
Transformer	4층 법원이 16일 오후 4층에서 열린 '4층 지법 법원 법원 법원에 참여하고 있었다. 이들은 "법원이 참여를 이용해주기 때문에 참여한 사람들이 참여하기 때문이라고 말했다. 그는 "법원과 마음을 이용해보고 있는 사람들과의 마음을 마련해 보기도 했다.
Transformer+Copy	법원이 16일 오후 자기 소통을 높이기 위해 마음을 높이고 있다. 법원은 16일 법원에서 법원에 참여하기 위해 스트레스를 이용하는 것을 마련했다. '요층'는 점심시간이 참여하고 있는 체질의 체질을 돌아보고 있었다.
MASS	법원 6층에 마련된 요가실에서 판사 4명과 법원 직원 40명은 점심시간을 이용해 법원이 마련한 '사상체질과 스트레스' 강좌에 참여했다. 판사 16명과 법관 40명이 점심시간으로 마련한 '사상체질와 스트레스' 강좌에 참석했고, 법원 직원 30명은 저녁 시간을 활용해 자신을 돌아보는 명상을 하고 있었다.
MASS+Copy	14일 오후 법원 6층에 마련된 요가실에서는 판사 4명이 스스로를 돌아보는 명상을 하고 있었다. 판사 16명과 법원 직원 40명은 점심시간을 이용해 법원이 마련한 '사상체질과 스트레스' 강좌에 참여했다. 14일 법원 6층에 마련한 요가실에서는 소송 당사자들과의 소통 능력을 높이기 위해 판사들도 자기관리를 한다.
MASS+Copy(Both)	판사 16명과 법원 직원 40명은 점심시간을 이용해 법원이 마련한 '사상체질과 스트레스' 강좌에 참여했다. 14일 오후 법원 6층에 마련된 요가실에서는 판사 4명이 스스로를 돌아보는 명상을 하고 소송 당사자들과의 소통 능력을 높이기 위해 판사들도 자기관리를 하고 있다.