

# PORTFOLIO

Jungbin Son, Machine Learning Engineer

---

Phone

+82 10-9242-0995

Email

[sonjbin@gmail.com](mailto:sonjbin@gmail.com)

# Jungbin Son



언어와 시각 정보를 연결하는 멀티모달 AI를 연구하는 손정빈입니다.  
문서 이미지에서 구조화된 정보를 추출하는 모델을 개발하고,  
이를 효율적으로 배포 및 서빙하였습니다.  
사람처럼 세상을 이해하는 AI를 만드는 것을 목표로 하고 있습니다.

## Carrier

23.10 - present Lomin (ML engineer)

## Education

21.09 - 23.09 KAIST 전산학부 석사  
17.03 - 21.08 KAIST 전산학부 학사  
14.03 - 17.02 창원 과학고등학교

## Web Site

Profile: <https://sonjbin.github.io/>  
GitHub: <https://github.com/sonjbin>

## Tech Stack

Development  
Python  
GitHub  
Hugging Face

Model Training  
PyTorch, PEFT (LoRA)  
DeepSpeed

Model Serving  
vLLM  
Triton Inference Server

# Publication

---

## Time-Aware Representation Learning for Time-Sensitive Question Answering

Jungbin Son\*, Alice Oh. *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP-Findings 2023, Short)*

- 기존 Question Answering (QA) 모델이 'after 2002', 'between 1998 and 1999'와 같은 시간 표현을 제대로 이해하지 못하는 문제를 해결하기 위해, 새로운 태스크와 합성 데이터를 기반으로 학습을 진행하였습니다. 제안한 Time-Context dependent Span Extraction (TCSE) 방식은 TimeQA 데이터셋에서 기존 baseline 대비 최대 8.5점의 F1-score 향상을 보였습니다.

## An Efficient and Effective Document Deduplication by Using Similarity-Based Clustering

Jungbin Son\*, Sunkyoung Kim, Minsoo Kim. *Korea Computer Congress 2021 (KCC 2021, 1577-1579.)*

- 한국어 뉴스 데이터 간 유사도를 측정하는 다양한 방식을 분석하고, DBSCAN 클러스터링을 기반으로 중복 제거 시스템을 설계하였습니다. 제안한 방법은 기존 알고리즘 대비 약 380배 빠른 처리 속도와 4% 높은 정확도를 달성하였습니다.

# Lomin

Document AI 기업 lomin에서  
ML Engineer로 수행한 프로젝트 내용입니다

---

Vision Language Modeling and Serving

---

Synthetic Image Generation

---

Context-Aware Document Classification

# Vision Language Modeling and Serving

무역서류의 정보 추출을 위한 Vision Language Model (VLM) 및 추론 파이프라인 개발

stack: PyTorch, DeepSpeed, vLLM, Triton inference Server

- Challenges
  - 대규모 Vision-Language Model 기반 범용 문서 이해 모델 개발
    - 문서 이해를 위한 대규모 VLM 개발: 한정적인 GPU 자원을 이용해 8B 이상의 모델 학습 필요.  
파라미터가 많은 모델을 GPU에 모두 로드할 수 없는 문제 발생.
    - 대규모 VLM 학습 시 문서 레이아웃 분석 등의 복잡한 문서 인식 태스크에 대한 성능이 낮음
  - 무역서류 정보 추출을 위한 도메인 특화 경량 Vision-Language Model 개발
    - 비정형 무역 문서는 다양한 레이아웃과 표현 방식으로 구성되어 있어, 기존 text-bbox 기반 모델 [1] 만으로는 정확한 정보 추출이 어려움
    - 적은 데이터를 이용하여 대규모 모델 학습 시 모델의 일반화 능력이 떨어지며 학습 데이터에 오버피팅 되는 현상 발생
    - Pseudo labeling된 데이터를 포함하여 학습 시 데이터의 노이즈가 모델에 반영되어 성능이 하락하는 현상 발생
    - Bounding box를 자연어 토큰으로 표현 시 box 하나당 10개 이상의 토큰이 필요해 추론 속도가 매우 느려지는 문제 발생
- Contributions
  - 대규모 모델 학습 최적화
    - Deepspeed Zero, low rank adaptation과 같은 모델 학습 최적화 기법을 적용하여 4개의 GPU에서 peak memory 15GB 미만으로 8B 모델 학습 성공
    - 모든 task를 한 번에 학습 시 어려운 task에 대한 성능이 떨어지는 문제 발생. 단순 텍스트 인식에서 구조화된 정보 추출 까지 점점 task의 난이도를 높여가는 curriculum learning을 구현하여 복잡한 문서 인식 task에서의 VLM 성능 개선

[1] "BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents". Hong et al., AAAI 2022

# Vision Language Modeling and Serving

무역서류의 정보 추출을 위한 Vision Language Model (VLM) 및 추론 파이프라인 개발

stack: PyTorch, DeepSpeed, vLLM, Triton inference Server

- Contributions

- **Instruction 학습 데이터 구축**

- OCR, key-value 추출, 테이블 추출로 구성된 instruction 기반 학습 데이터 생성 및 각 task를 구분하는 special token 기반의 instruction prompt 설계
    - 복잡한 문서인식 task를 학습해야 했으나 라벨링된 학습 데이터가 1000장 미만으로 매우 부족한 상황 발생.  
부족한 학습 데이터를 보완하기 위해 pre-trained VLM을 활용하여 노이즈가 포함된 pseudo label 자동 생성 시스템 구축

- **모델 Pre-training & Fine-tuning**

- 라벨 신뢰도에 따라 샘플 가중치를 조정하는 weighted training 전략으로 pseudo-labeled 데이터의 노이즈에 robust한 학습 구현
    - 추론 속도 개선을 위해 bbox 표현 방식 실험: 자연어 토큰 vs 좌표 표현을 위한 special token 두 가지 방식으로 모델 성능 비교 후 special token 방식을 적용하여 토큰 수 60% 이상 절감.

- **모델 Deploying & Serving**

- vLLM OpenAI-compatible 서버를 활용한 고속 추론 API 및 서빙 파이프라인 구축

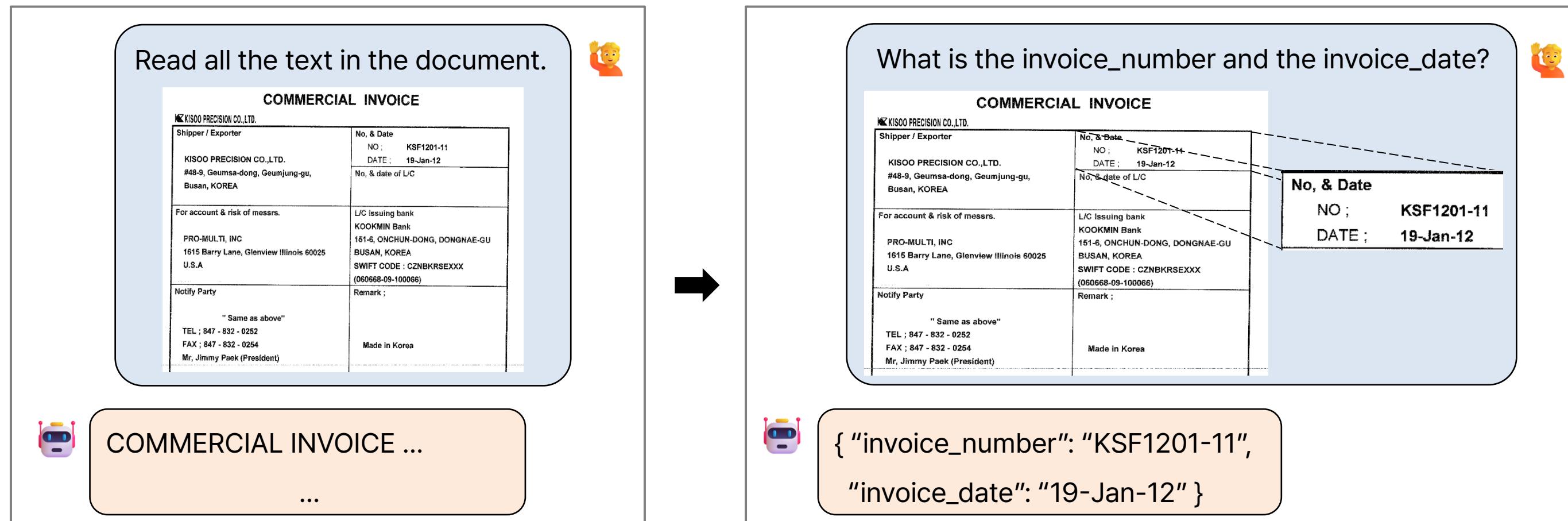
- Results

- 12개 회사 무역 문서에서 key-value 및 table 인식률 96% 달성 및 문서당 평균 추론 시간 2초로 실시간 처리 가능 수준 확보

# Vision Language Modeling and Serving

## Key-value, table 인식을 위한 VLM 학습 구성

- Pre-training
  - 대규모 문서 이미지에서 OCR task를 통해 텍스트 인식 및 좌표 표현 방식을 학습
- Fine-tuning
  - 문서 내 특정 key-value 쌍 및 테이블 데이터를 구조화된 JSON 형태로 반환하는 방식으로 추가 학습 진행



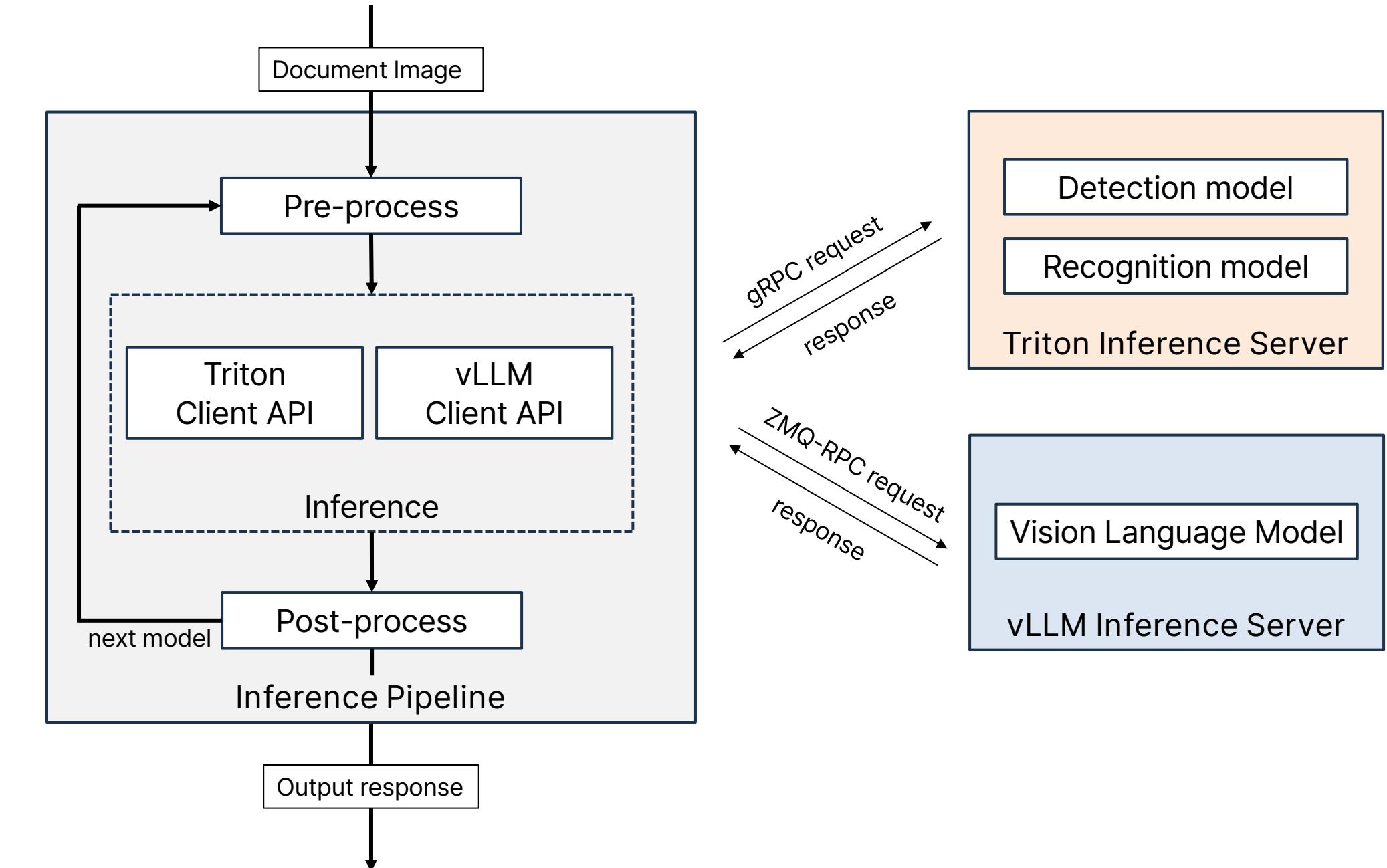
Pre-training

Fine-tuning

# Vision Language Modeling and Serving

## On-premise 환경에서 사용 가능한 VLM 추론 파이프라인 구축

- 추론 파이프라인의 모델 추론 순서
  - 텍스트 감지: Detection model
  - 텍스트 인식: Recognition model
  - Key-value, table 정보 추출: Vision Language Model
- 각 모델 추론의 work flow
  - Pre-process: 입력 데이터를 모델 입력 형식에 맞게 변환
  - Inference: Inference 서버에 추론 요청 전송
  - Post-process: 결과를 구조화된 최종 결과 포맷으로 변환



# Vision Language Modeling and Serving

## 무역서류의 정보 추출을 위한 추론 파이프라인 결과 시각화 예시

추출된 각 항목에 대한 class와 신뢰도 수치를 시각화 한 결과

# COMMERCIAL INVOICE

TD-CI-DON:COMMERCIAL INVOICE

score: 1.00

KISOO PRECISION CO.,LTD.

Shipper / Exporter		No, & Date	
<b>KISOO PRECISION CO.,LTD.</b>	NO :	<b>KSF1203-06</b>	
TD-CI-SPR:KISOO PRECISION CO.,LTD.	DATE :	TD-CI-CID:23-Mar-12	
score: 1.00		score: 1.00	
<b>#48-9, Geumsa-dong, Geumjung-gu,</b>		TD-CI-CID:23-Mar-12	
TD-CI-SHA:#48-9, Geumsa-dong, Geumjung-gu,		score: 1.00	
score: 1.00		score: 1.00	
Busan, KOREA		score: 1.00	
For account & risk of messrs.		L/C Issuing bank	
<b>PRO-MULTI, INC</b>		TD-CI-MKL/C Issuing bank	
TD-CI-ARM:PRO-MULTI, INC		score: 1.00	
1615 Barry Lane, Glenview Illinois 60025		KOORMIN Bank	
TD-CI-ARAD:1615 Barry Lane, Glenview Illinois 60025		score: 1.00	
U.S.A		151-6, ONCHUN-DONG, DONGNAE-GU	
score: 1.00		BUSAN, KOREA	
score: 1.00		SWIFT CODE : CZNBKRSEXXX	
score: 1.00		(060668-09-100066)	
Notify Party		Remark ;	
<b>" Same as above"</b>			
TD-CI-NP: "Same as above"			
score: 1.00			
TEL : 847 - 832 - 0252			
TD-CI-NFT:847-832-0252			
score: 1.00			
FAX : 847 - 832 - 0254			
TD-CI-NFF:847 - 832 - 0254			
score: 1.00			
<b>Mr. Jimmy Paek (President)</b>		Made in Korea	
TD-CI-NFAT:Mr. Jimmy Paek (President)			
score: 0.99			
Port of loading		Final destination	
<b>BUSAN-PORT</b>		<b>CHICAGO, IL</b>	
TD-CI-POL:BUSAN-PORT		TD-CI-FD:CHICAGO, IL	
score: 1.00		score: 1.00	
Carrier		Sailing on about	

## Key-value 인식 결과

Mark and number of	Description of goods	Quantity (each)	Unit price (US\$)	Amount (US\$)
<b>FG25008-6</b> TD-CI-PAR:FG25008-6 score: 1.00	<b>PIVOT-SECTION</b> TD-CI-TNM:PIVOT-SECTION score: 1.00	2,592 TD-CI-OTY:2,592 score: 1.00	11.52 TD-CI-UNP:11.52 score: 1.00	29,859.84 TD-CI-AMT:29,859.84 score: 1.00
<b>FG20060-XA</b> TD-CI-PAR:FG20060-XA score: 1.00	<b>FLANGE GEAR CASE</b> TD-CI-TNM:FLANGE GEAR CASE score: 1.00	720 TD-CI-OTY:720 score: 1.00	5.10 TD-CI-UNP:5.10 score: 1.00	3,672.00 TD-CI-AMT:3,672.00 score: 1.00
<b>FG20061-XA</b> TD-CI-PAR:FG20061-XA score: 1.00	<b>FLANGE GEAR CASE</b> TD-CI-TNM:FLANGE GEAR CASE score: 1.00	828 TD-CI-OTY:828 score: 1.00	4.98 TD-CI-UNP:4.98 score: 1.00	4,123.44 TD-CI-AMT:4,123.44 score: 1.00
<b>20178-0300</b> TD-CI-PAR:20178-0300 score: 1.00	<b>FLANGE GEAR CASE</b> TD-CI-TNM:FLANGE GEAR CASE score: 1.00	809 TD-CI-OTY:809 score: 1.00	4.20 TD-CI-UNP:4.20 score: 1.00	3,397.80 TD-CI-AMT:3,397.80 score: 1.00
<b>20160-0800</b> TD-CI-PAR:20160-0800 score: 1.00	<b>UNDER HOLDER</b> TD-CI-TNM:UNDER HOLDER score: 1.00	1,260 TD-CI-OTY:1,260 score: 1.00	2.55 TD-CI-UNP:2.55 score: 1.00	3,213.00 TD-CI-AMT:3,213.00 score: 1.00
<b>ES-20128-3</b> TD-CI-PAR:ES-20128-3 score: 1.00	<b>END STRG SHAFT</b> TD-CI-TNM:END STRG SHAFT score: 1.00	1,000 TD-CI-OTY:1,000 score: 1.00	4.22 TD-CI-UNP:4.22 score: 1.00	4,220.00 TD-CI-AMT:4,220.00 score: 1.00
<b>14</b> TD-CI-TTP:14 score: 1.00	<b>Case</b> TD-CI-TUP:Case score: 1.00	Total FOB TD-CI-FOB score: 1.00	7,209 TD-CI-TO:7,209 score: 1.00	48,486.08 TD-CI-TTA:48,486.08 score: 1.00
<b>Packing List</b>				
C/T		Signed by		
N/W :				
G/W :	10,200 KG TD-CI-GW:TD-CHTUG:KG score: 1.00			
M/M :	11.51(CBM) TD-CI-MD:CHTUM:CBM score: 1.00			
				<b>PRESIDENT ; Jungsoo,Jeon</b>

Table 인식 결과

# Synthetic Image Generation

## 한글 및 한자 문서 이미지의 인식을 위한 OCR 성능 개선

- Challenges

- 기존 OCR 모델은 세로쓰기 문서(종서) 및 한자 인식률이 낮음
- 종서는 대부분의 글자를 띠어쓰기 없이 쓰지만 대부분의 현대 문서에 띠어쓰기가 있기 때문에 세로로  
붙어있더라도 detection 모델이 box를 분리하도록 예측하는 문제 있음
- 기존의 합성 데이터 생성 방식 [2]은 종서를 지원하지 않으며, 생성된 문서의 레이아웃과  
실제 고서 형태의 차이 존재

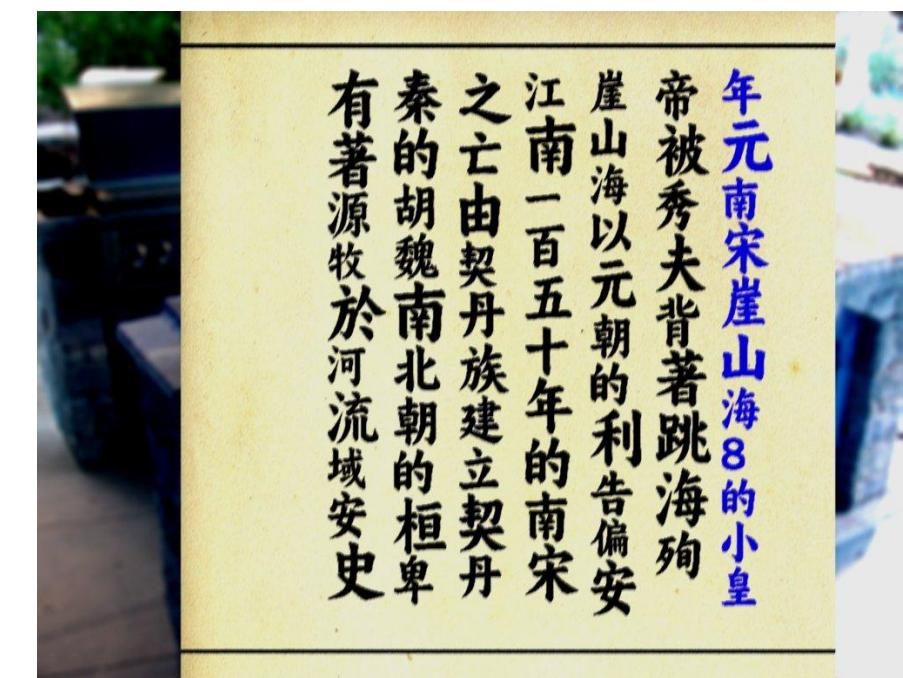
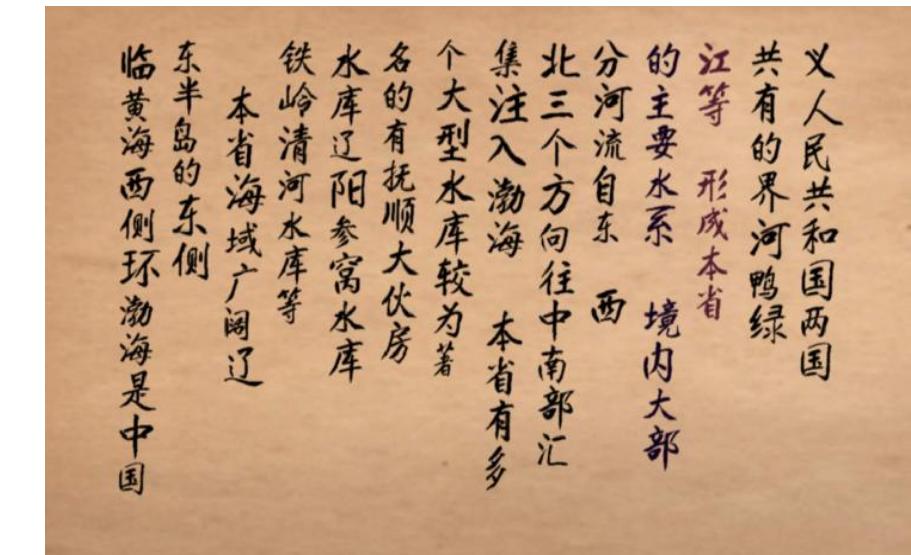
- Contributions

- 세로쓰기 형태 지원 및 다양한 문서 스타일 반영한 합성 데이터 생성 파이프라인 구현
- 한자 인식률 향상을 위한 data augmentation 기법 설계
  - Detection 문제를 해결하기 위해 랜덤하게 여러 단어를 붙여 이미지 생성
  - 실제 고서에서 나타나는 위 아래 가로선과 노이즈를 랜덤하게 추가
- 생성 데이터를 활용한 텍스트 검출 및 인식 모델 학습
- 실제 고서 기반의 한자 테스트 데이터 수집 및 성능 평가

- Results

- 가로 및 세로쓰기 형태를 포함한 8종의 테스트셋 기준 평균 17% 인식률 향상

stack: PyTorch, Data generation, Multi-lingual model



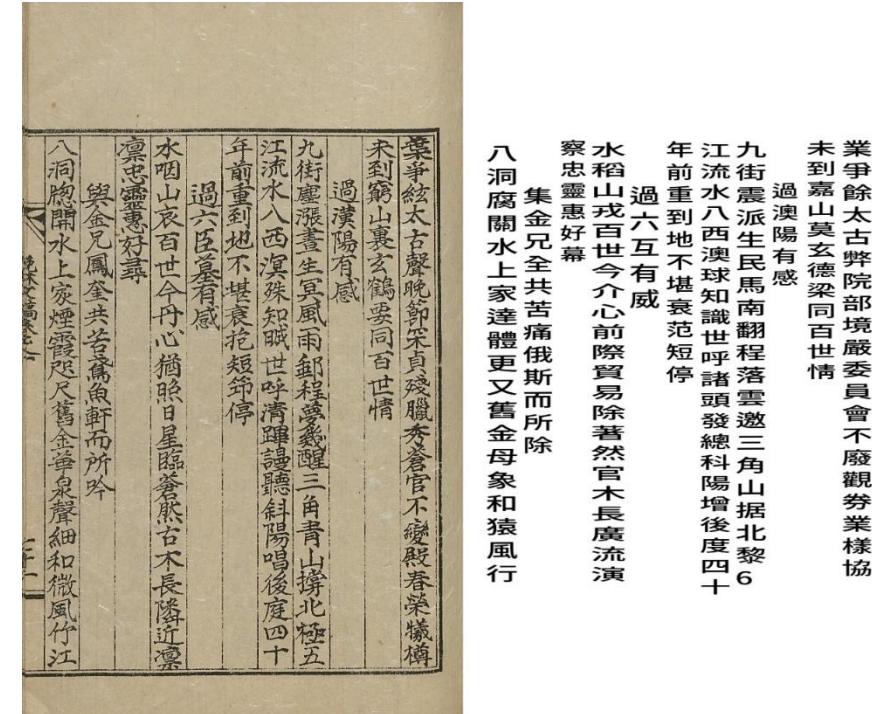
생성된 고서 이미지 예시

# Synthetic Image Generation

- 학습 데이터 생성
  - 레이아웃, 문서 내용(corpus), 서체, 배경 이미지를 랜덤 조합해 총 4,000장의 한자/한글 문서 이미지 생성

- 테스트 데이터 수집
  - 문서 구조, 언어 구성, 서체 유형 기준 총 8종의 한자 테스트셋 수집
    - 텍스트 방향:** 횡서(가로쓰기) / 종서(세로쓰기)
    - 언어 구성:** 한자 / 한자-한글 혼합
    - 서체 유형:** 인쇄체(인출본, 활자본) / 필기체(해서, 행서)

- 모델 성능 평가 방식
  - 정답 word box와 IOU 기준으로 가장 유사한 예측 box 매칭
  - 매칭된 box 내에서 character-level F1-score 계산
  - 전체 예측에 대한 평균 score로 모델 성능 평가



한자 인식 결과 예시

# Context-Aware Document Image Classification

정밀한 문서 분류를 위한, 문서의 context를 이해하는 Image classification 모델 개발

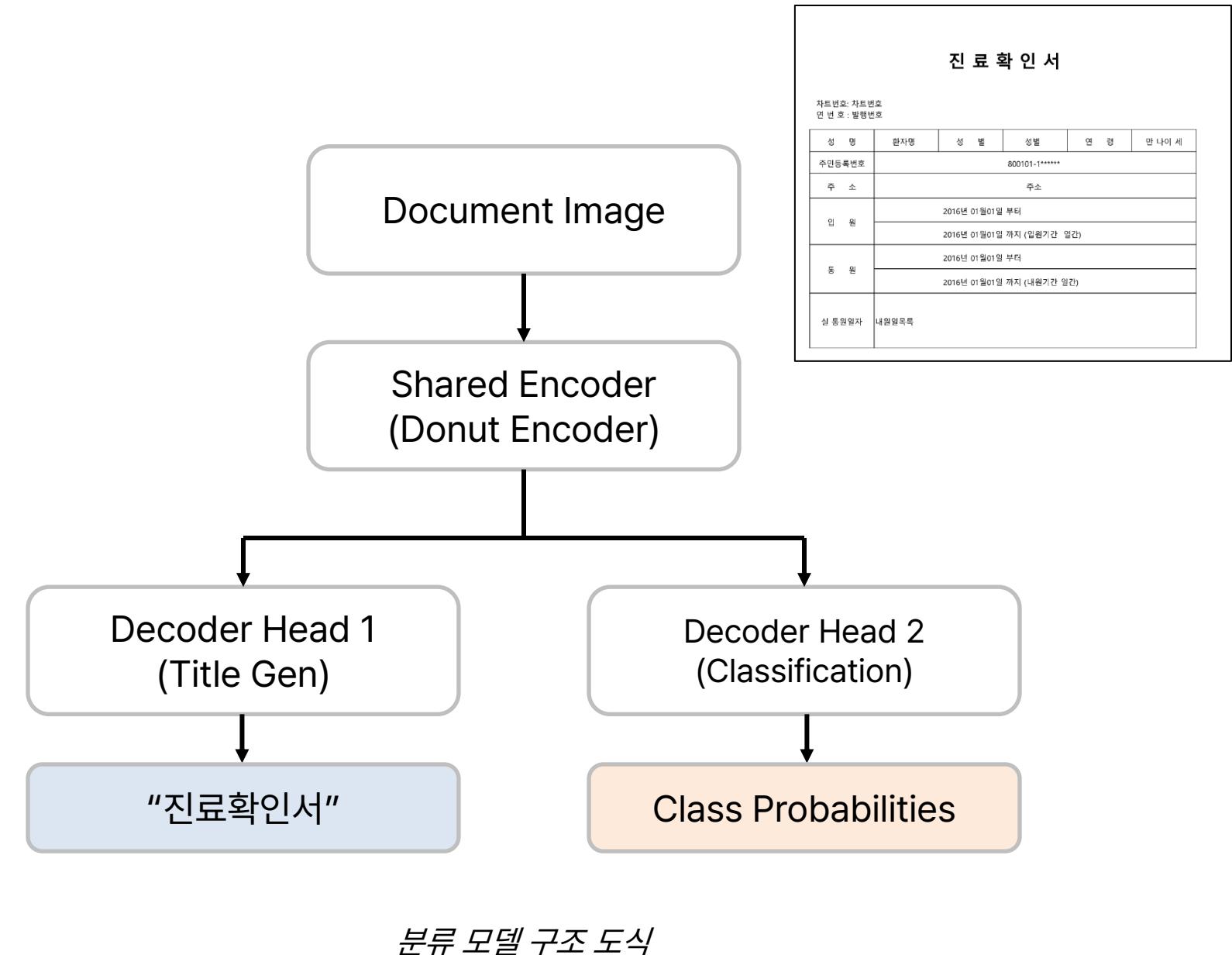
stack: PyTorch, Multi-task learning

- Challenges
  - 입원 확인서, 퇴원 확인서와 같은 문서 제목만 다른 유사 레이아웃 문서 간 분류 성능 저하
- Contributions
  - 제목은 문서의 주요 내용을 담고 있으므로 제목을 찾을 수 있다면 문서의 맥락을 잘 파악할 수 있을 것이라 생각하여 이를 중심으로 모델 개선 진행
  - Donut [2] 모델을 기반으로 제목 생성을 위한 3가지 학습 방식 테스트
    1. 문서 분류 토큰과 제목을 순차적으로 생성하도록 학습 → Decoder가 문서 분류시 제목을 참고할 수 없는 문제
    2. 제목 생성 이후 문서 분류 토큰을 순차적으로 생성하도록 학습 → 분류 토큰이 항상 동일한 위치에 생성되지 않는 문제
    3. two-head encoder-decoder 구조로 변경하여 multi-task 학습 가능하도록 설계 → 두개의 head를 사용함으로써 각 task에 대한 간섭 없이 효과적으로 문서의 맥락을 파악할 수 있는 분류 모델 개발 성공
- Results
  - 유사한 형태가 많은 의료 문서(입퇴원확인서, 입원확인서 등)의 분류성능 f1-score 최대 18% 까지 향상

# Context-Aware Document Image Classification

## Context-Aware Training 방식 설계

- 모델 구조
  - **Donut base:** image encoder + text decoder
    - Output: [Title\_1], ..., [Title\_n], [SEP], [CLS\_LABEL], [PAD], ...
    - 두 task 간 간섭으로 인해 분류 성능 저하
  - **Context-aware donut:** image encoder + (2 \* text decoder)
    - Output\_title: [Title\_1], ..., [Title\_n], [PAD], ...
    - Output\_cls: [CLS\_LABEL], [PAD], ...
    - 각 task를 효율적으로 학습 했으며 개선된 성능 보임
- 모델 학습
  - 제안한 모델의 두 decoder head의 학습 task:
    - Decoder 1: 문서 제목 생성 (Title Generation)
    - Decoder 2: 문서 분류 (Classification)
  - Loss 구성: 각 task에 대한 cross entropy loss의 weighted sum 계산



분류 모델 구조 도식

# Publication

학사, 석사 과정 중 연구한 주제로 작성한 논문입니다

---

Time-Aware Representation Learning for Time-Sensitive Question Answering

---

An Efficient and Effective Document Deduplication by Using Similarity-Based Clustering

---

# Time-Aware Representation Learning for Time-Sensitive Question Answering

Jungbin Son\*, Alice Oh. EMNLP 2023

## 연구 목적

- Question Answering (QA) 모델이 'after 2002', 'between 1998 and 1999' 와 같은 시간 표현과 숫자와의 관계를 인식하지 못함
- 시간 관련 태스크를 수행하도록 모델을 학습 시킬 충분한 데이터셋이 아직 없으며, 효율적인 학습 방법 필요

## 1. Time-Context dependent Span Extraction (TCSE) task 정의

Q: Who **worked in the Salvation Army before 1995?**

A: Harry

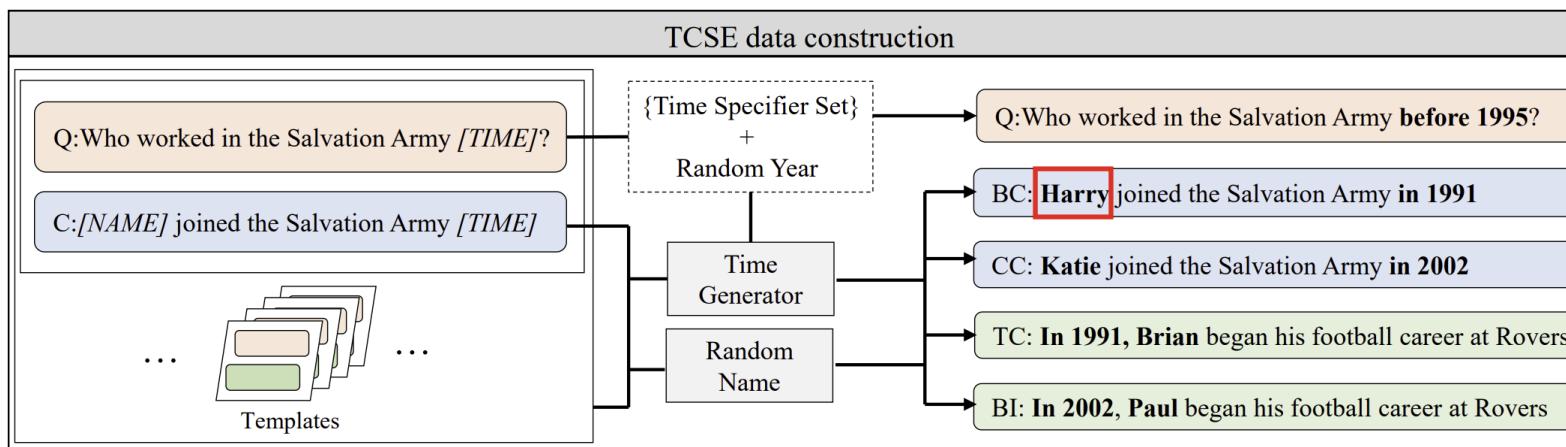
Context Time	Correct	Incorrect
Correct	<p><u>Harry joined the Salvation Army in 1991</u></p> <p>Both Correct (BC)</p>	<p><i>In 1991</i>, Brian <b>began his football career at Rovers</b></p> <p>Time Correct (TC)</p>
Incorrect	<p>Katie <b>joined the Salvation Army in 2002</b></p> <p>Context Correct (CC)</p>	<p><i>In 2002</i>, Paul <b>began his football career at Rovers</b></p> <p>Both Incorrect (BI)</p>

- QA 모델이 시간 표현을 이해하지 못하는 것은 기존의 QA 데이터셋이 다양한 시간표현을 학습할 수 있을 만큼 충분한 데이터가 없기 때문이라고 판단
- 모델의 시간 표현 학습을 위해 시간 제약 조건을 포함한 질문과 시간, 맥락의 관점에서 구분된 네 개의 문장으로 구성되는 질의 응답 태스크 (TCSE) 제안

# Time-Aware Representation Learning for Time-Sensitive Question Answering

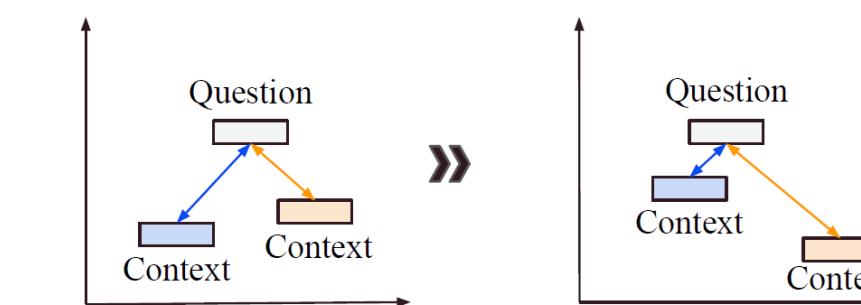
Jungbin Son\*, Alice Oh. EMNLP 2023

## 2. TCSE 데이터 생성 프레임워크 및 데이터셋 구축



## 3. Contrastive Learning을 활용한 시간 표현 학습

$$L_{CLR} = \sum_{v_q, v_c, Y \in S} w_p Y * \exp \left( \text{dist}(v_q, v_c) \right) + w_n (1 - Y) * \exp \left( 1 - \text{dist}(v_q, v_c) \right)$$

 $v_c, v_q$ : embedding of question and context, respectively $w_p, w_n$ : weight of positive and negative sample, respectively**Y: label (positive = 1, negative = 0)**

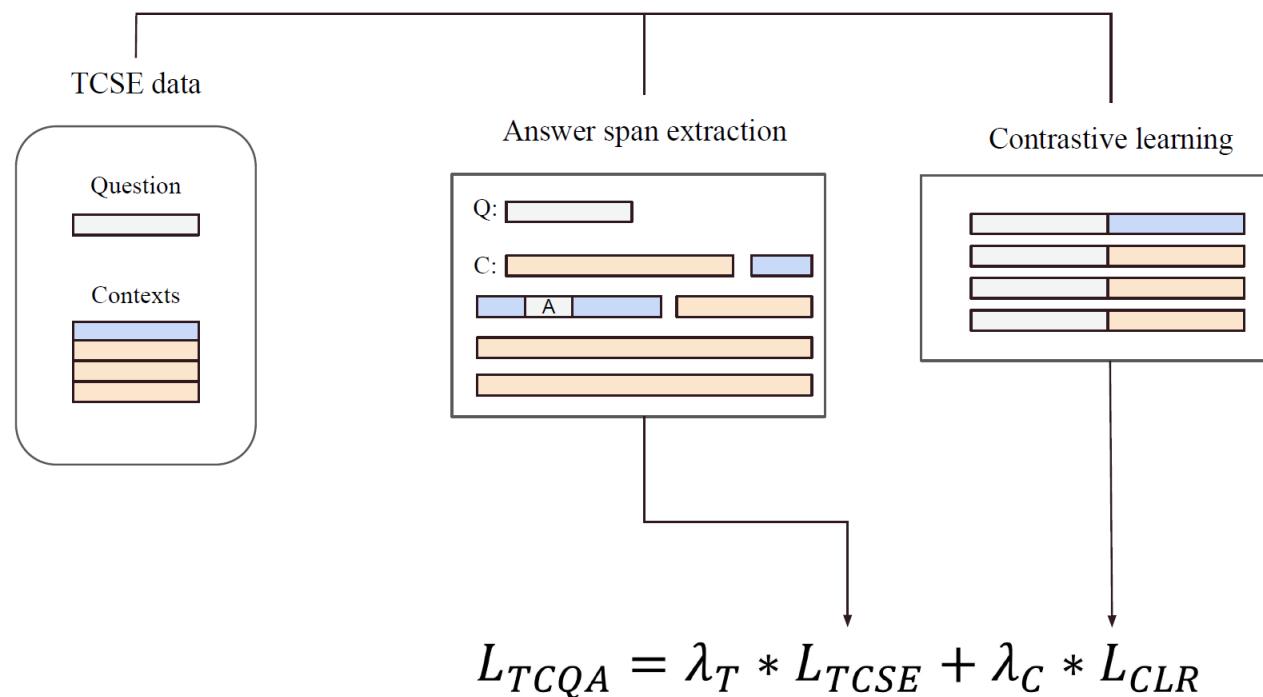
- Wikipedia 문서에서 시간 표현이 포함된 문장 수집
- 생성 모델인 T5 모델을 이용한 질문 생성으로 Question-context 템플릿 생성
- 템플릿에 시간 표현을 삽입하여 TCSE 데이터 자동 생성 시스템을 구축

- 벡터 공간 상에서 맥락, 시간표현이 맞지 않는 문장 사이의 거리를 멀게 학습시키기 위해 contrastive learning 활용
- TCSE 데이터 샘플에서 맥락, 시간표현이 모두 맞는 positive pair 하나와 negative pair 세 개를 데이터 쌍으로 구성해 contrastive loss 학습

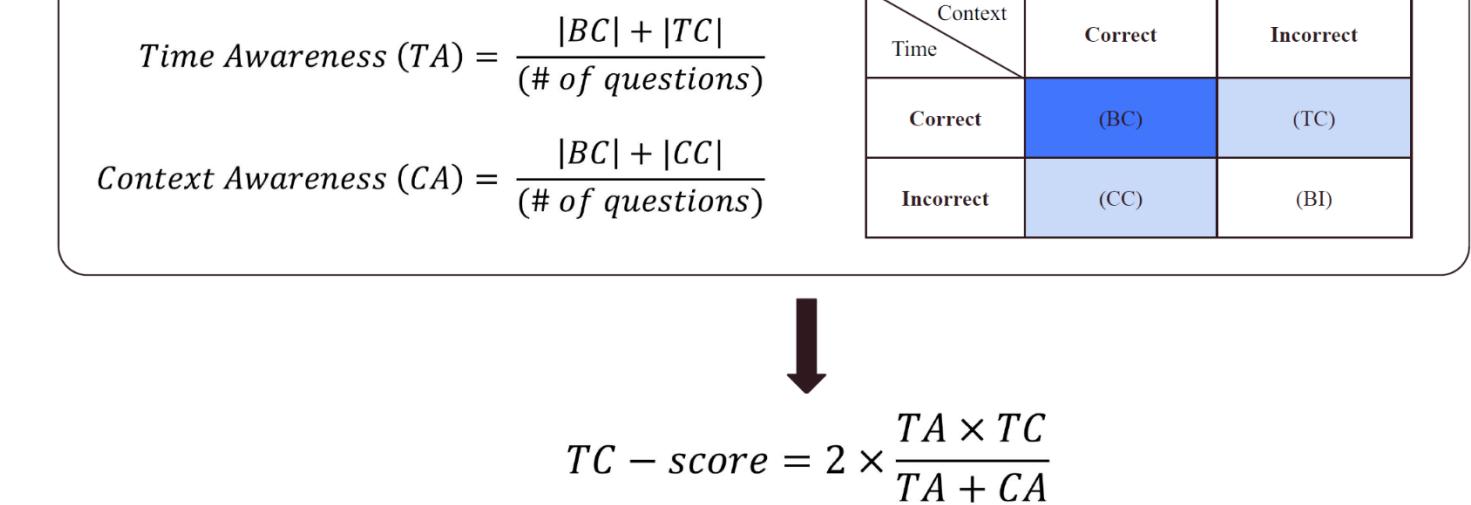
# Time-Aware Representation Learning for Time-Sensitive Question Answering

Jungbin Son\*, Alice Oh. EMNLP 2023

## 4. Time-Context aware Question Answering (TCQA) Framework 제안



## 5. 평가 지표 Time-Context awareness Score (TC-score) 제안



- Span extraction 과 contrastive learning을 multi-task learning으로 학습하는 학습 프레임워크 제안

- TCSE 태스크에서 네 문장이 각각 시간, 맥락이 대응되는지에 따라 분류됨을 이용한 새로운 평가 지표 제시
- 맞는 시간 범위에서 정답을 추출한 문제의 비율과 맞는 맥락 상에서 정답을 추출한 문제의 비율 값의 조화평균을 계산하여 QA 모델의 맥락과 시간 이해도 정량적 평가

# Time-Aware Representation Learning for Time-Sensitive Question Answering

Jungbin Son\*, Alice Oh. EMNLP 2023

## 1) TimeQA Result

Model	$\text{BERT}_{\text{base}}$		$\text{RoBERTa}_{\text{base}}$		$\text{ALBERT}_{\text{base}}$		$\text{Bigbird}_{\text{RoBERTa}}$	
Metric	EM	F1	EM	F1	EM	F1	EM	F1
Baseline	19.95	26.25	29.89	38.5	24.66	33.5	44.61	53.56
+TCQA	<b>25.63</b>	<b>34.75</b>	<b>30.86</b>	<b>39.03</b>	<b>27.36</b>	<b>35.48</b>	<b>46.31</b>	<b>54.26</b>

→ TimeQA 데이터셋에서 모두 Baseline 성능 뛰어넘음

## 2) Qualitative Analysis

Question	Passage	Answer (BigBird)	Answer (BigBird+TCQA)
What position did John Pope take between Sep 1831 and Nov 1833?	... He served as a member of the Kentucky Senate from 1825 to 1829 , and ... ... From 1829 to 1835 , he served as the Governor of Arkansas Territory . ...	member of the Kentucky Senate (X)	Governor of Arkansas Territory (O)

→ Test set에서 시간 범위를 잘 판단하여 답변

## 3) TC-score Validation

	Time Awareness (TA)	Context Awareness (CA)	TC - score
BigBird FT on NQ*	51.48	<b>88.78</b>	65.16
BigBird FT on TimeQA	<b>67.96</b>	79.32	<b>73.21</b>

→ 제안하는 방법을 통해 언어 모델의 Time awareness와 Context awareness를 효과적으로 측정

# An Efficient and Effective Document Deduplication by Using Similarity-Based Clustering

Jungbin Son\*, Sunkyoung Kim, Minsoo Kim. KCC 2021

## 연구 목적

- 기존의 문서 중복 제거 방식은 문서 쌍 단위로 비교하여 중복의 정도를 판별하여 제거하기에 많은 계산량 필요
- 이는 대용량 문서 텍스트에 적용하기에 어려운 한계점이 있으므로 해결하고자 함

## 1. 뉴스 기사 패턴 분석 및 전 처리 작업

- 중복되는 내용으로 판단되는 뉴스 기사의 패턴(문장 재배치, 단어 추가 및 제거 등) 분석
- 형태소 단위로 토큰화 하여 주요 형태소만 사용하는 방식으로 데이터 전 처리 진행

## 2. 문서 클러스터링 알고리즘 구현

- 문서 벡터화 후 벡터 공간 상에서 유사도 기반 클러스터링 진행
- 문서 임베딩 벡터: TF-IDF, Doc2Vec, SentenceBERT을 각각 사용한 문서 임베딩 비교하여 뉴스 기사 클러스터링 정확도가 가장 높은 TF-IDF로 선정
- 클러스터링 알고리즘: 문서 간 거리를 코사인 유사도로 사용하였으며, 다섯 가지의 클러스터링 알고리즘으로 실험하여 수행 시간과 정확도를 고려하여 가장 효율적인 DBSCAN 이용

# An Efficient and Effective Document Deduplication by Using Similarity-Based Clustering

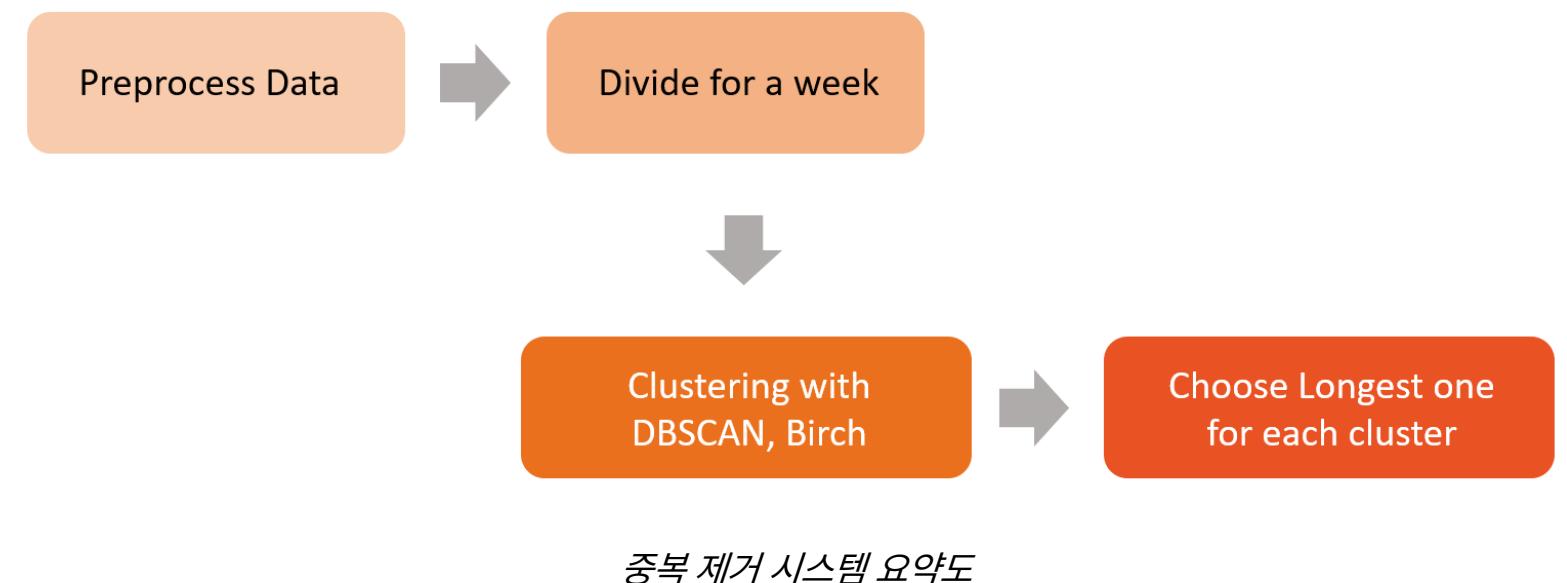
Jungbin Son\*, Sunkyoung Kim, Minsoo Kim. KCC 2021

## 3. 뉴스 데이터 중복 제거 시스템 구축

- 전체 뉴스 데이터를 일주일 단위로 나누어 각 시간 범위마다 클러스터링 진행
- 각 클러스터에서 한 개의 문서를 랜덤으로 추출함으로써 중복 제거 수행

## 4. 문서 클러스터링 평가 데이터셋 구축

- 중복 제거 알고리즘의 성능 비교를 위한 평가 데이터셋 구축
- 네이버 헤드라인 뉴스가 동일한 주제의 뉴스 그룹을 제공함을 이용, 크롤링을 통해  
9279 개의 뉴스 기사 및 431 개의 클러스터로 구성된 데이터셋 수집
- 레이블과 클러스터링 결과 간 사이의 Normalized Mutual Information (NMI) 계산



## 연구 결과

- 기존 알고리즘 대비 중복 제거 속도 약 380배 향상
- 클러스터링 평가 지표 NMI 4% 이상 향상

# PORTFOLIO

---

Phone

+82 10-9242-0995

Email

[sonjbin@gmail.com](mailto:sonjbin@gmail.com)