KCC 2022

Sequence-to-sequence 모델에 기반한 한국어 대화 요약

유호선 / 전북대학교 / rhsblossom@naver.com
이건희 / 전북대학교 / dkghszkfhs@jbnu.ac.kr
나승훈 / 전북대학교 / nash@jbnu.ac.kr

01

연구 배경 및 목표

- **대화(dialogue) 요약(summarization)의 필요성**
- 비대면 의사소통의 중요성 대두, 대규모 대화 텍스트 축적 중
- Ex) 콜센터 등에서 직원과 민원인의 상담 대화 요약, 여러 명의 화자가 참여한 업무 미팅 요약 등

- **대화 요약의 어려움**
- 대화 즉 구어체의 특성상 적은 단어로 구성됨
- 2명 이상의 화자가 상호 소통하는 특성

- **기존 시도들**
- 요약 decoder에 청자와 화자 정보를 라벨링

Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts. Chih-Wen Goo, Yun-Nung Chen. 2018.

- 대화를 문서 형태로 변형한 뒤 요약 진행

Restructuring Conversations using Discourse Relations for Zero-shot Abstractive Dialogue Summarization. Prakhar Ganesh, Saket Dingliwal. 2020.

-> Label이 충분한 데이터셋에서 좋은 성능 기대 가능

- 한국어 대화 요약을 위해
- Seqeunece-to-sequence 모델 기반 요약문 생성
- 사전 학습 모델 활용 pre-train & finetuning

02

관련 연구

비지도 학습 기반 대화 요약

RepSum: Unsupervised Dialogue Summarization based on Replacement Strategy

Xiyan Fu, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Changlong Sun, Zhenglu Yang

- RepSum 방법론 -

- 뛰어난 요약문은 원문 대화를 대체할 수 있는 특성을 가짐!
- 요약문의 성능을 확인하기 위해 보조 태스크(auxiliary task)를 수행할 때 성능을 측정
- 추출(extractive) 요약과 생성(abstractive) 요약 적용
- 다른 비지도 학습 기반 요약 기법에 비해 좋은 성능을 보임

요약 과제에 알맞은 pre-train objective 실험

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu

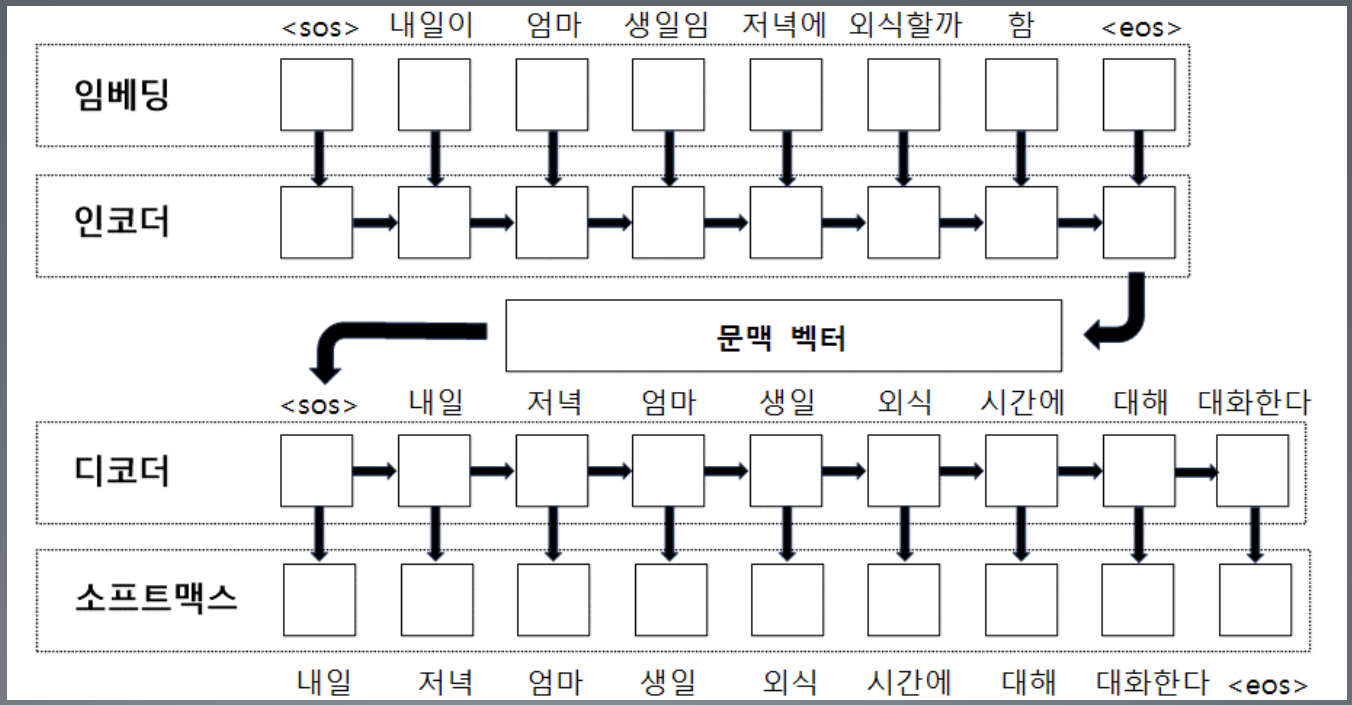
- PEGASUS 방법론 -

- 가정: pre-training objective와 downstream task가 밀접한 관계일 수록 성능 향상
- -> 일반성을 포기하고, abstractive summarization에 적합한 pre-training objective 제시

03

모델 구조

예시: LSTM을 사용한 모델 구조



$$h_t = EncoderRNN(e(x_t),\ h_{t-1})$$
$$s_t = EncoderRNN(d(y_t),\ s_{t-1})$$

$$\hat{y} = softmax(\ s_t)$$

04

실험 및 평가

학습 데이터

P1: "내일이 #@MASK# 엄마 생일임."
P2: "저녁에 외식할까 함."
P1: "시간 ㅇㅋ?"
P2: "강의만 제때 끝나면 저녁에 시간돼요"
P3: "저도 가능해요"

summary: 가족 구성원들이 내일 저녁 엄마 생일 외식 시간에 대해 대화한다

데이터 전처리

1. Json 형태의 데이터에서 uttererance 추출
2. 특수문자 제거, 반복 자음 이모티콘 통일(ex. ㅋㅋㅋㅋ, ㅋㅋㅋ -> ㅋㅋ)
3. 형태소 단위 tokenizing
4. 단어 사전 구성
   - 2회 이상 출현 토큰, 소스 텍스트: 72,190개, 타겟 텍스트: 34,616개

05

실험 및 평가

모델 평가 - ROUGE score

		ROUGE-1	ROUGE-2	ROUGE-L
LSTM	Recall	0.47	0.11	0.39
	precision	0.22	0.06	0.19
	F-Measure	0.28	0.08	0.25
Transformer (base)	Recall	0.47	0.10	0.47
	precision	0.33	0.07	0.33
	F-Measure	0.39	0.08	0.39
Transformer (KoBART)	Recall	0.47	0.26	0.41
	precision	0.46	0.24	0.40
	F-Measure	0.45	0.24	0.40

06

요약문 생성

요약문 생성 샘플 - good case

원문	아근데 내가 바이탈 잴 기회가 없자너 / 웅웅 / 연습좀 하둘려고 너한테 할거야 ㅋㅋ / ㅋㅋ들오면 한명씩해ㅋㅋ / 아 저앗어
요약문 (정답)	바이탈을 잴 기회가 없어서 연습을 해야 하는데 다른 사람들 오면 한 명씩 연습해보라고 이야기한다
요약문 (생성)	바이탈 잴 기회가 없어서 연습 좀 하두려고 한다
원문	할머니가 젓국인가 그걸 끓였는데 따뜻한 젓갈냄새 엄청심해 / 아 진짜 너무짜증나 / 할머니방으로 피신하면 안대나 / 그거 냄새 진짜 괴로워 / 온집이 다그레 / 환풍기 틀구 해따 / 나도 냄새에 민감해서 이해해
요약문 (정답)	할머니가 젓국을 끓이셨는데 냄새가 지독하다
요약문 (생성)	할머니가 끓인 젓국에서 따뜻한 젓갈냄새가 너무 심해서 할머니방으로 피신하면 안되나싶다

07

요약문 생성

요약문 생성 샘플 - bad case

원문	오늘도 팀장한테 잡힐뻔 배아파서 살음 ㅌㅋㅋㅌ/ ㄷ ㄷ 미쳣나봐 싫다 진짜 / 나랑 백순대 먹으려고함
요약문 (정답)	오늘도 팀장이 나랑 부평 백순대를 먹으려고 하였는데 배가 아파서 붙잡히지 않았다
요약문 (생성)	오늘도 팀장에게 잡혀서 백순대를 먹으려고 한다
원문	주임님임니꺄 / ㅋㅋ자요 따뜻할때놔뒀는데 다식엇겟네 둥글레차 / 새건데요 ㅎㅎ 꺄오 잘먹고 집갑니댱 / 주말잘보내융 / 녜녜 포항 조심해서 갔다와용 / 즐거운 주말 보내쎄요
요약문 (정답)	둥굴레 차 새거 라니까 잘 먹고 간다고 인사한다
요약문 (생성)	따뜻할 때 둥굴레 차를 놔뒀는데 다 식었다

08

요약문 생성

실제 대화 요약문 생성

[오전 9:32] ㅋㅋㅋㅋ ᄆ 다들 여름엔 볼수있음?
[오전 9:32] 이모티콘
[오전 9:32] 조아요~!
[오전 9:34] 여름이 기대되는구만 ㅋㅋㅋㅋ ᄆ
[오전 10:01] 날씨더우니까 계곡갑시다><
[오전 10:54] 계곡 너~~~~~~무 좋다!!!!!!!!
[오전 10:55] 꺄악 재밌겠다!!!!!!!!!!
[오전 11:00] 오 대박이네
[오전 11:08] 8월 주말에 가야되나용?ㅋㅋ
[오전 11:08] 우웅 ㅠㅜ 나 7월은 반기 끝나고 부가세도 잇고 바쁘다
[오전 11:08] 차라리 8월이 좋을듯!!!
[오전 11:09] 8월 좋구만
[오전 11:09] 8월 주말 이상무
[오전 11:09] 8월 주말로 추진해보도록하죠
[오후 12:51] 저두 7월 좀 지나봐야알듯용 ㅠㅠ

summary: '8월 주말에 계곡에 가기로 했다'