KUBIG CONFERENCE Poster Template(AI가 평가해주는 플러팅)

18th Haewon Jung, 19th Seunghyun Shim, Jiwon Bae, Hayoun Chung
KUBIG(Korea University Data Science & AI Society), Korea University
(ttom0505, chenny0811, com0901226, jchung02)@korea.ac.kr

KUBIG DATA SCIENCE & AI 6th CONFERENCE

Motivation & Key Idea

그동안 AI가 출력하는 메시지를 사람이 평가해왔다
⇅
사람의 메시지를 AI가 평가하면 어떨까?

기존 챗봇에 기능을 추가해, 사람이 메시지를 입력하면 AI도 메시지를 출력하여 대화를 이어나가는 동시에 AI가 인간을 평가한 점수도 동시에 출력하도록 한다

1. Topic
- 20대가 제일 관심있어 하는 주제인 연애를 주제로 선정
- 사용자가 AI챗봇과 대화하면, AI가 대화하며 상대방에게 느끼는 호감도를 계산하여 함께 출력하는 챗봇을 구성. 이는 기존 연애 관련 챗봇들과 차별화되는 점

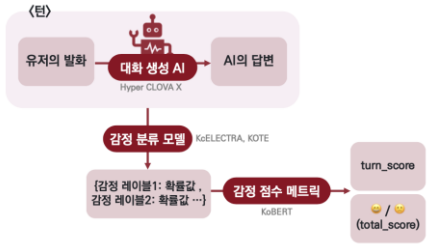
2. 핵심 요소
- 자연스러운 대화가 가능한 챗봇 구현
- 호감도 점수 산출하는 메트릭 개발 후 호감도 점수 계산 후 출력
- 대화 및 호감도 출력값이 납득 가능해야함

3. 학습 요구 사항
- 일상 연인간 대화 데이터를 활용해 친근한 텍스트 내역 학습
- 각 대화 턴마다 호감도 점수 계산 메트릭 학습
- 대화의 맥락을 고려하도록 학습 필요

4. 웹 앱으로 구현
- 학습한 챗봇을 사용자들이 직접 체험할 수 있도록 웹 앱 서비스로 구현 필요

Approach & Model

- **시스템 아키텍처:**



1. 대화형 챗봇
- CLOVA Studio 튜닝 '대화' 기능
- 훈련 데이터셋: 채팅 데이터, ['Text', 'Completion'] 형태로 입력
- 각 대화 시나리오 당
  1) 발화 문장을 [누적 대화 – 마지막 답변]으로 재구조화
  2) P02의 마지막 발화를 기준으로 시나리오 종료
  3) 매 시나리오에 대해 동일한 작업

	Text	Completion
1	P01의 1번째 발화 문장	P02의 1번째 발화 문장
2	P01의 1번째 발화 문장, P02의 1번째 발화 문장, P01의 2번째 발화 문장	P02의 2번째 발화 문장
···	···	···
n	n−1 행까지의 발화 누적(sep = ' ')	P02의 n번째 발화

2. 호감도 점수 산출
[호감도 점수 = AI의 답변 문장에 대한 감정 레이블 각각 확률 * 감정 별 가중치]

- 모델: KcELECTRA-base + 채팅 데이터, 감정 레이블
  ⇒ 채팅/대화 텍스트에 특화된 감정 레이블, 각각의 확률값 도출
- 감정 레이블 별 가중치: KoBERT 워드 임베딩을 활용한 '설렘', '관심', '호감', '사랑'과 각 감정 라벨 간의 코사인 유사도 점수의 평균 ⇒ 긍/부정 0-1 사이 스케일링
- Turn Score: AI의 답변 문장에 대한 Σ(감정 라벨 각각 확률 * 감정 별 가중치)
- Total Score: 모든 AI의 답변 문장에 대한 Σ(감정 라벨 각각 확률 * 감정 별 가중치)
  ※ 최신 턴에 큰 가중치 부여 (현재 턴 수 / 총 대화 수) 대화 수 = 턴 수 * 2

Datasets

1. SNS 데이터 고도화(AiHub)
- 2020~2021년 한국어 일상대화 데이터 셋
- 연인 간 대화 상황 추출을 위해 화자가 남녀가 번갈아 나오고 '개인 및 관계' 주제인 대화 상황 추출
- 총 5만여 개의 대화 행을 추출

2. KOTE(Korean Online That-gul Emotions)
- 다양한 플랫폼에서 수집한 5만여 개의 댓글을 43개 감정 레이블로 분류한 데이터셋
- 각 감정 레이블의 확률값과 가중치 도출
- 문장 그 자체가 아닌 대화 참여자가 느끼는 감정을 내포
- 챗봇 대화에 감정 라벨링 하기 위해 사용함.

Results

- **Streamlit을 통한 웹 앱 구현**

1) 사용자가 채팅창 화면에 메시지를 입력해 서버에 전송
2-1) 채팅메시지를 생성하는 첫 모델의 결과 출력
2-2) 호감도 점수를 계산하는 두,세 번째 모델의 결과 출력(turn score, total score)