

Week1 학습정리

🕒 생성일	@2022년 2월 21일 오후 12:28
👤 생성자	☺ 손지아

Week1-1. Orientation

참고자료 읽기

1. NLP review paper : <https://arxiv.org/pdf/1708.02709.pdf> ★★★★★
 - a. 2017년 버전 번역 : [https://ratsgo.github.io/natural language processing/2017/08/16/deepNLP/](https://ratsgo.github.io/natural%20language%20processing/2017/08/16/deepNLP/)
 - b. 2022년 동영상 : <https://www.youtube.com/watch?v=B3qd3LgsTRY>
2. Pytorch vs TensorFlow in 2022 : <http://www.assemblyai.com/blog/pytorch-vs-tensorflow-in-2022/> ★★★★★

Week1-2 NLU (Natural Language Understanding)

목표 : 내가 풀어야할 문제가 어떤 task인지 구별할 수 있고
benchmark 데이터 셋 및 모델을 알 수 있다.

NLU란?

Natural Language Understanding

- Understanding
 - Syntactic : 문법적으로 옳은 문장인지 구분할 수 있는가?
 - Semantic : 문장의 의미를 아는가?
 - 의미(추상적) : 감정 분석(긍정/부정), 문장 간 유사도, 의도 파악, 질문에 대답, 추론

모델이 문법을 잘 맞추고 문장이나 대화의 의미를 잘 알고 있다면 언어를 이해하고 있다고 할 수 있다.

GLUE & SQuAD benchmark

세계에서 가장 공신력 있는 NLU 대회 (benchmark)

과제	벤치마크	예제	
감정 분석 Sentiment Analysis	SST, IMDB, NSMC	입력	문장: 신만이 이 영화를 용서할수 있다
		출력	긍정 · 부정
유사도 예측 Similarity Prediction	STS, MRPC, QQP, PAWS-X	입력	문장1: 파이썬과 자바 중 뭐부터 배워야 하나요? 문장2: Java나 Python 중 하나를 배워야 한다면, 뭐부터 시작해야 할까요?
		출력	유사 · 비유사
자연어 추론 Natural Language Inference	SNLI, MNLI, XNLI, aNLI	입력	전제: 회색 개가 숲에서 쓰러진 나무를 핥고 있다. 가설: 개가 밖에 있다.
		출력	참 · 거짓 · 모름
언어적 용인 가능성 Linguistic Acceptability	CoLA	입력	문장: 그는 한 번도 프랑스에 갔다.
		출력	가능 · 불가능
기계 독해 Reading Comprehension	SQuAD, RACE, MS MARCO, KorQuAD	입력	지문: 시카고 대학의 학자들은 경제학, 사회학, 법학 분석, 문학 비평, 신학, 행동주의 정치학 등 다양한 학문 발전에 중요한 역할을 담당해왔다. [...] 이 대학은 미국 최대 대학 출판사인 시카고 대학 프레스(University Press of Chicago Press)의 본거지이기도 하다. 오는 2020년 완공될 버락 오바마 대통령 센터에는 오바마 대통령 도서관과 오바마 재단 사무소가 세워질 예정이다. 질문: 버락 오바마 대통령 센터의 완공 예정연도는?
		출력	2020년
의도 분류 Intent Classification	ATIS, SNIPS, AskUbuntu	입력	질문: 현재 위치 주변에 있는 맛집 두 곳만 알려줘
		출력	장소 검색 (식당, 가까운, 평점 좋은, 2)

GLUE

- 문법 (CoLA) : 이진 분류 문제
- 감성 분석 (SST) : 1~5 중 선택
- 문장 유사성 (STS, MRPC, QQP)
- 추론 (MNLI, RTE)
- 언급 대상 추론 (WNLI) : Winograd NLI
 - 대명사, 지시사가 무엇을 가리키는지 명확히 하는 문제

SQuAD

























- QA : 지문에서 질문에 대한 답을 찾는다. 시작 인덱스와 끝 인덱스를 반환

SuperGLUE : GLUE 보다 어려움 NLU 태스크

- paperwithcode benchmark : <https://paperswithcode.com/dataset/superglue>

Benchmarks

[Edit](#)

Trend	Task	Dataset Variant	Best Model	Paper	Code
	Natural Language Inference	RTE	DeBERTa-1.5B		
	Common Sense Reasoning	ReCoRD	DeBERTa-1.5B		
	Question Answering	BoolQ	T5-11B		
	Question Answering	COPA	DeBERTa-Ensemble		
	Question Answering	MultiRC	DeBERTa-1.5B		
	Word Sense Disambiguation	Words in Context	COSINE + Transductive Learning		
	Natural Language Inference	CommitmentBank	DeBERTa-1.5B		
	Coreference Resolution	WSC	GPT-3 175B		

[Collapse benchmarks](#)

- SuperGLUE 논문의 태스크 예시

Table 2: Development set examples from the tasks in SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. Underlined text is specially marked in the input. Text in a monospaced font represents the expected model output.

BoolQ	Passage: <i>Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</i> Question: <i>is barq's root beer a pepsi product</i> Answer: No
CB	Text: <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i> Hypothesis: <i>they are setting a trend</i> Entailment: Unknown
COPA	Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i> Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i> Correct Alternative: 1
MultiRC	Paragraph: <i>Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week</i> Question: <i>Did Susan's sick friend recover?</i> Candidate answers: <i>Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)</i>
ReCoRD	Paragraph: <i>(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the <u>US</u> commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the <u>State Electoral Commission</u> show. It was the fifth such vote on statehood. "Today, we the people of <u>Puerto Rico</u> are sending a strong and clear message to the <u>US Congress</u> ... and to the world ... claiming our equal rights as <u>American</u> citizens, <u>Puerto Rico</u> Gov. <u>Ricardo Rossello</u> said in a news release. @highlight <u>Puerto Rico</u> voted Sunday in favor of <u>US</u> statehood</i> Query <i>For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency</i> Correct Entities: US
RTE	Text: <i>Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.</i> Hypothesis: <i>Christopher Reeve had an accident.</i> Entailment: False
WiC	Context 1: <i>Room and <u>board</u>.</i> Context 2: <i>He nailed <u>boards</u> across the windows.</i> Sense match: False
WSC	Text: <i>Mark told <u>Pete</u> many lies about himself, which Pete included in his book. <u>He</u> should have been more truthful.</i> Coreference: False

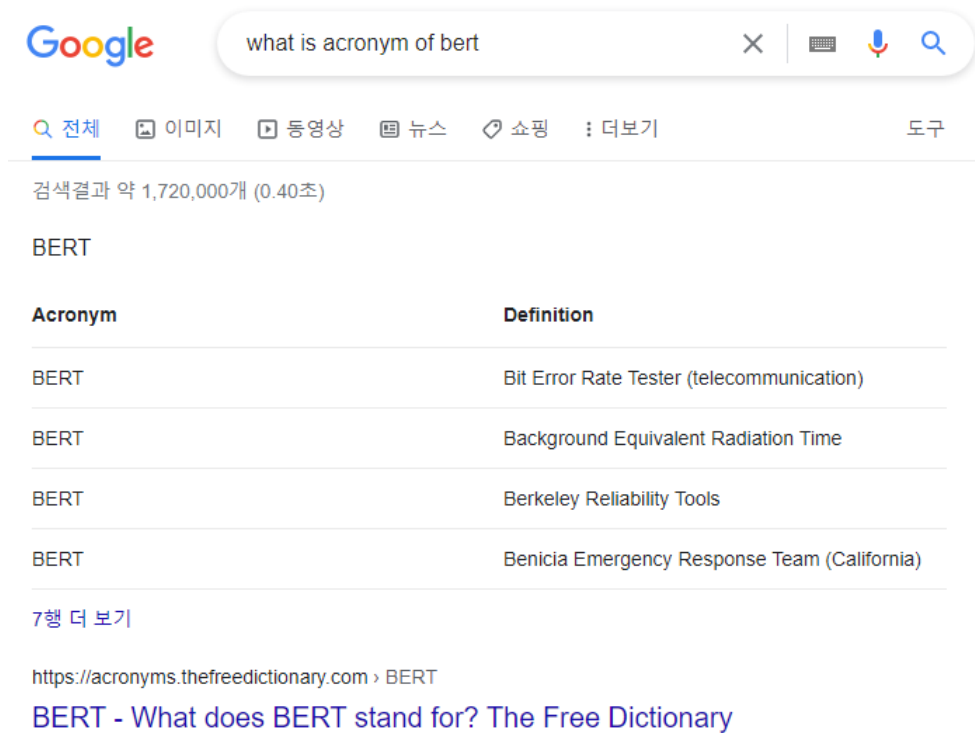
SQuAD2.0

- 더 어려운 QA : 표에서 답 찾기

NLP task in real life

- 문법 (CoLA) → 자동 문법 교정
 - 예: Grammarly
- 감성 분석 (SST) → 상품 & 서비스 리뷰 데이터 공부정 판별
 - 마케팅 팀에서 리뷰 분석 혹은 리뷰 태깅에 사용
- 문장 유사성 (paraphrase) (STS, MRPC, QQP) → 유사 문서 클러스터링
 - 예 : Quora Question Pair Competition - Quora의 유사한 의도를 가진 질문 묶는 대회
- 추론 (MNLI, RTE) → 자동 내부 및 회계 감사
 - 문장이 법률과 참/거짓/중립 여부 판별
- 언급 대상 추론 (WNLI) → QA, 요약, 번역 등의 성능 향상을 위해 필수 요소
 - 요약, 번역, QA 서비스의 베이스, 코어 태스크

- QA (SQuAD) → 검색 시스템 스니펫



NLP task SOTA model

GLUE SOTA model

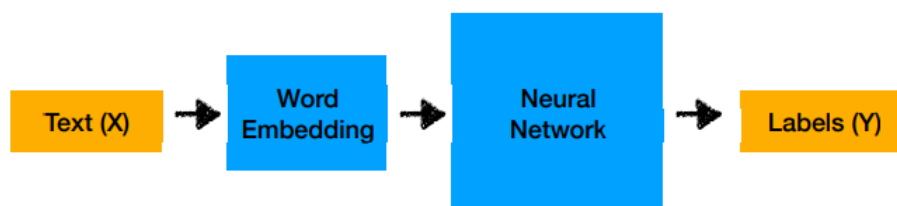
- 문법 : biLSTM, BERT(RoBERTa)
- 감성 분석 : BERT(RoBERTa)
- 문장 유사성 : BERT(RoBERTa)
- 추론 : T5, BERT(ALBERT, DeBERTa)
- 언급 대상 추론 (WNLI) : BERT(SpanBERT, RoBERTa)

SQuAD SOTA model

- QA : T5, Bert(Bigbird), XLNet
 - QA는 대형 모델이 성능이 좋음

Applications

- 텍스트 분류 : 이탈 고객 예측, 상품 카테고리 분류, 위험 판별



- 텍스트 군집화 : 유사 제품군 군집화, 유사 키워드 생성
 - Vectorize : distance (K-means)
 - 예 : 제품 리뷰의 벡터 간 거리를 구해서 군집화 가능

Reference

- [한글 블로그 - NLP task 및 데이터 소개](#) ★★★★★
- [GLUE](#) ★★★★★
- [한국어 - KLUE](#) ★★★★★
- [SQuAD2.0](#) ★
- [한국어 - KorQuAD2.0](#) ★
- [Paperswithcode](#) ★★★★★

Week1-3 NLG (Natural Language Generation)

목표 : NLG의 세부 task를 이해할 수 있다.

NLG란?

언어를 '생성'한다는 말은 무엇일까?

- 주어진 정보(text, video, ...)를 기반으로 정보 축약, 보강, 재구성해서 문장을 만드는 것
 - 축약 : "많은 데이터 중에서 핵심만 뽑아"
 - 보강 : "부족한 정보를 추가해줘"
 - 재구성 : "기존 정보의 형태를 변형해줘", "기존 정보로 추론하고 답해줘"

Text Abbreviation(축약)

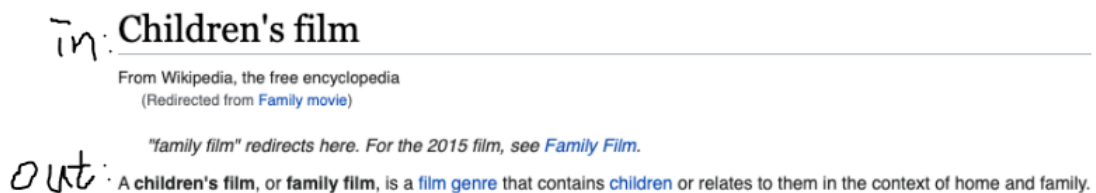
긴 문장에서 핵심 추출

- Summarization : 정보량이 늘어서 요약이 중요해져서 현업에서 수요가 높음. 뉴스 요약에서 사용
 - Abstractive : 없던 단어로 요약 생성하기에 어려운 문제
 - Extractive : 문장 간 랭킹 매기고 추출 요약
- Question generation : 문서를 보고 답을 찾을 수 있는 질문 생성. 데이터 증강에 사용하거나 online 학습 도구로 사용
- Distractor generation : 오지선다에서 오답을 생성, 데이터 증강에 사용

Text Expansion(보강)

부족한 정보 추가

- Short text expansion : 정보를 추가해 데이터 길이를 늘리기, 짧은 제목으로 내용 생성



- Topic to essay generation
 - 예 : [입력] "가족", "장남감", "크리스마스" → [출력] "케빈은 크리스마스에 혼자 집에서 ..."

Text Rewrite(재구성)

기존 문서 형태를 변형하거나 추론을 통해 정답 생성

- Style Transfer
 - 문체 긍정 ↔ 부정 변환

i got a bagel breakfast sandwich and it was delicious! → i got a bagel breakfast sandwich and it was disgusting!

- 응용 : 데이터 증강, 말투 변화(오픈 데이터셋 없음/파파고 사투리 번역)
- Dialogue Generation
 - 페르소나를 갖은 참여자의 대화를 생성
 - 대용량 모델에서 application으로 많이 제공
 - 응용 : 챗봇

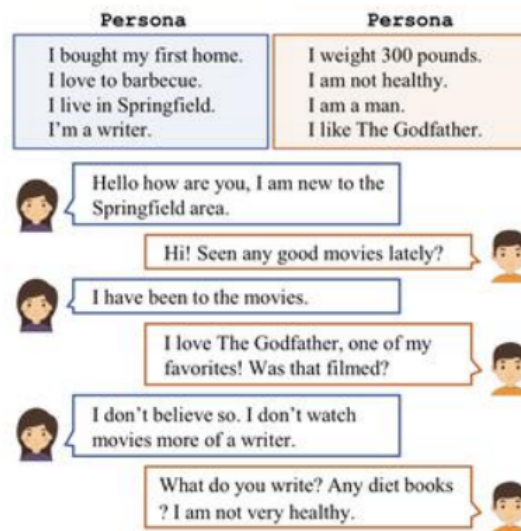


Figure 1: A clipped dialogue from PERSONA-CHAT.

Models

NLG SOTA 모델을 구성하는 주요 아키텍처

- Encoder decoder
 - RNN Seq2seq
 - Transformer
- Copy and pointing
- Gan (Generative Adversarial Network)
- Memory network
- GNN (Graph Neural Network)
- External Knowledge

Reference

- 논문 - 최신 NLG 리뷰 논문 ★★★
- [영문 블로그 - seq2seq 모델 구조](<https://jalammar.github.io/illustrated-transformer/>) ★★
- [영문 블로그 - GAN 구조 설명](https://developers.google.com/machine-learning/gan/gan_structure) ★★
- 문서 요약 서비스 - PLEX