Stanceformer: Target-Aware Transformer for Stance Detection

Krishna Garg **Cornelia Caragea**
kgarg8@uic.edu cornelia@uic.edu
Computer Science
University of Illinois Chicago

Venue: EMNLP 2024

발제자: HUMANE Lab 석사과정생 이다현

2024-12-13

Introduction

- Stance Detection은 특정 타겟에 대한 텍스트의 입장을 파악하는 작업
- 타겟은 텍스트에서 표현된 입장의 맥락과 주제를 정의
- 여러 연구들은 모델들이 예측할 때 타겟을 간과하는 경향이 있다고 지적(Yuan et al. 2022; Kaushal et al. 2021)
  - 모델의 성능을 향상시키기 위해서는 타겟을 인식하도록 보장하는 것이 중요

Text: a woman ?? wanting to be equal to a man ???! what montrosity is this
Target: feminist movement
Stance: Against

- 기존 연구는 데이터 증강, 외부 지식 베이스, human-like reasoning, 대조 학습 등의 방법으로 접근
- BiLSTM 기반 모델들에 대한 target-specific attention을 향상한 연구 존재
  (Xu et al., 2018; Du et al., 2017; Augenstein et al., 2016)
- 이 연구는 **transformer 모델에서 target-specific attention**을 향상시키는 메커니즘을 개발

Contribution

1. Stanceformer: Stance Detection 작업을 위해 설계된 transformer 모델
   - 타겟에 대한 attention를 향상시키기 위한 Target Awareness Matrix를 제안
2. Stanceformer가 최신 transformer 모델들을 효과적으로 대체할 수 있다는 것을 실험적으로 입증
3. Stance Detection을 위해 LLM들을 파인튜닝한 첫 번째 연구
   네 개의 stance detection 데이터셋에 걸쳐 Llama-2-chat 모델들을 Fine-Tuning
4. 감성분석과 같은 다른 도메인에서도 방법의 일반화 능력을 입증

Task Description

학습 데이터

- $D_{train} = \{(x_i, t_i, y_i)\}_{i=1}^{n}$
- 텍스트 시퀀스

$$x_i = [x_{1i}, x_{2i}, \ldots, x_{li}]$$

- 타겟 시퀀스

$$t_i = [t_{1i}, t_{2i}, \ldots, t_{pi}]$$

- stance 레이블

$$y_i \in \{1, \ldots, c\}$$

테스트 데이터

- $D_{test} = \{(x_j, t_j)\}_{j=1}^{m}$
- 입력 형태

$$[\mathrm{CLS}]x_i[\mathrm{SEP}]t_i[\mathrm{SEP}]$$

- 전체 시퀀스 길이

$$\mathrm{seq} = 1 + l + 1 + p + 1$$

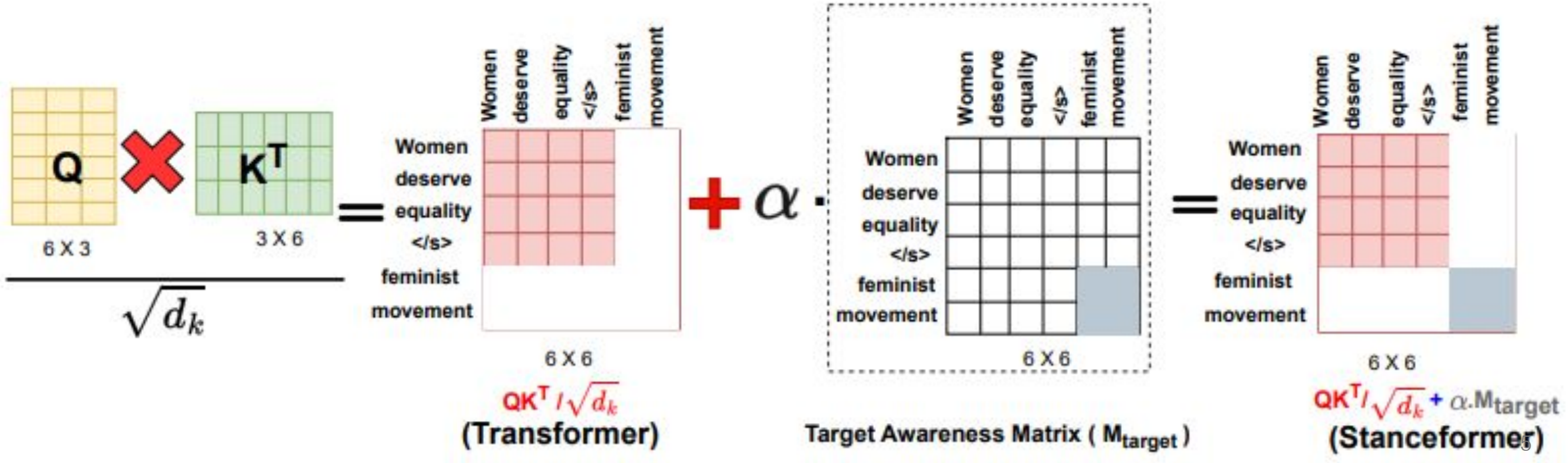
x_j와 t_j가 주어졌을 때 스탠스 레이블 $y_j \in \{1, \cdots, c\}$를 예측

Target-Awareness Matrix

- Self-Attention: 시퀀스 내 토큰의 중요도를 가중치로 계산 $\boxed{[\mathrm{CLS}]x_i[\mathrm{SEP}]t_i[\mathrm{SEP}]}$

$$SA = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}}) * V \quad \Rightarrow \quad SA' = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}} + \alpha M_{target}) * V$$

- Target Awareness Matrix: $M_{target} = (p \times p)$



Fine-tuning with Stanceformer

- 기존 Transformer 기반 모델에 Stanceformer를 통합할 수 있는 방식을 제안
- Transformer 모델의 학습된 가중치에 Target Awareness Matrix (M_{target})를 추가
- 학습 데이터에 대해 모델을 Fine-tuning하여, target 토큰에 대한 attention을 우선시하도록 함
- 일관성 보장을 위해 테스트 과정에서도 동일한 Target Awareness Matrix를 적용

Experimental Setup

Datasets

- SemEval-2016(Mohammad et al., 2016)
  - Twitter 데이터셋으로, 5개의 target 포함, 스탠스 레이블은 찬성, 반대, 중립
- COVID-19 (Glandt et al., 2021)
  - COVID-19 관련 트윗을 기반으로 구축된 데이터셋, 레이블은 찬성, 반대, 중립
- Pstance(Li et al, 2021)
  - 미국 내 세 명의 주요 정치 인물에 대한 개인들의 입장으로 레이블은 찬성 또는 반대
- VAST(Allaway and McKeown, 2020)
  - 대규모 데이터셋으로 다양한 주제를 포괄, 레이블은 찬성, 반대, 중립
  - VAST-zero-shot 버전은 Zero-shot Stande Detection에 맞게 특별히 구성된 하위 집합

Dataset	#Train	#Val	#Test	Targets
SemEval-2016	2,160	359	1,080	Atheism, Feminist Movement, Hillary Clinton, Legalization of Abortion
COVID-19	4,533	800	800	Face Masks, Fauci, Stay at Home Orders, School Closures
P-Stance	17,224	2,193	2,157	Joe Biden, Bernie Sanders, Donald Trump
VAST-zero-shot*	13,477	1,019	1,460	drug addict, gun, constitutional right, etc.

Experimental Setup

Baseline: 전체 데이터셋 실험

- BERT-based
  - 각 데이터셋의 이전 연구를 따라 각 데이터셋에 맞는 BERT의 변형을 사용
  - SemEval-2016과 PStance 데이터셋
    - BERTweet (Nguyen et al., 2020) - 트윗 데이터를 기반으로 학습
  - Covid-19 데이터셋
    - Covid-Twitter-BERT (Müller et al., 2020) COVID-19 관련 트윗 분석에 최적화
  - WS-BERT(He et al., 2022) Wikipedia에서 추출된 배경 정보를 통합
  - PT-HCL (Liang et al., 2022a) Zero-shot 스탠스 탐지에 특화된 모델로, 대조 학습(contrastive learning)을 활용
- non-BERT-based models
  - BiLSTM, CNN, TAN과 같은 BERT 기반이 아닌 모델들을 활용
  - 토픽 그룹화 어텐션 기반 방법인 TGA-Net과도 비교
  - +) VAST zero-shot 데이터셋에 대해서는 BiCond, CrossNet, SEKT, TPDG, TOAD도 추가
- LLM models
  - Llama-2-7b-chat
  - Llama-2-13b-chat

  ————————————
  - gpt-3.5-turbo-0613

Experimental Setup

Evaluation

- Macro F1을 사용
- 모든 수치는 3번의 독립적인 실행의 평균
- LLM의 경우, 한 번의 실행에 대한 결과만을 보고

Results

Full Dataset Settings

- 모든 데이터셋에서 베이스라인 보다 Stanceformer가 일관되게 나은 성능
- 각 대상별로 봤을 때도 성능 향상
- 대부분의 LLM은 BERT 기반 모델보다 낮은 성능
- GPT-3.5 > Llama-2 13B > 7B
- Stanceformer 변형은 Fine-tuned LLM에서도 최대 7% 성능 향상

	SemEval-2016					Covid19					PStance			
	AT	FM	HC	LA	Avg.	mask	fauci	home	school	Avg.	trump	biden	sanders	Avg.
BiCE†	64.88	57.93	58.81	60.86	57.23	56.70	63.00	64.50	54.80	59.75	77.15	77.69	71.24	75.36
CNN-based†	66.76	58.83	57.12	65.45	58.31	59.90	61.20	52.10	52.70	56.48	76.80	77.22	71.40	75.14
TAN‡	59.33	55.77	65.38	63.72	59.56	54.60	54.70	53.60	53.40	54.08	77.10	77.64	71.60	75.45
CrossNet	-	-	-	-	-	-	-	-	-	66.16	48.60	47.10	40.80	45.50
BERT‡	68.67	61.66	62.34	58.60	59.09	-	-	-	-	68.71	79.19	76.02	73.59	76.27
TGA-Net	-	-	-	-	-	-	-	-	-	69.09	-	-	-	77.66
BERT	65.19	55.95	63.01	61.08	61.31	71.13	72.52	77.60	61.77	70.76	79.81	79.34	76.61	78.59
-» Stanceformer	64.86	57.49	64.70	62.48	62.38	71.40	72.36	78.66	64.19	71.65	79.87	81.13	75.65	78.88
BERT-variant	68.15	60.06	65.77	62.93	64.23	79.23	81.25	83.16	**85.32**	82.24	80.92	81.24	75.91	79.36
-» Stanceformer	69.99	61.84	66.65	65.56	66.01	81.49	83.43	**87.49**	80.23	83.16	82.75	81.44	77.57	80.59
WS-BERT	70.38	63.20	71.33	62.99	66.98	82.59	82.48	84.53	81.09	82.67	84.97	82.86	79.97	82.60
-» Stanceformer	**72.01**	64.41	73.39	63.96	68.44	**85.10**	**83.79**	85.44	81.86	**84.05**	**85.35**	**83.96**	**80.57**	**83.30**
Closed-source LLM														
GPT-3.5 [0-shot]	24.92	69.41	73.27	57.94	56.38	76.90	73.03	72.81	50.96	68.42	79.80	79.65	77.77	79.07
Open-source LLM														
Llama-2-7b-chat [0-shot]	17.34	48.37	53.09	36.67	38.87	43.84	38.92	31.26	26.25	35.06	67.33	68.38	69.03	68.25
Llama-2-7b-chat-finetune	44.49	44.56	56.79	45.42	47.81	63.84	62.99	57.07	60.65	61.14	72.00	67.96	65.57	68.51
-» Stanceformer	49.13	48.40	55.11	40.51	48.29	68.00	73.25	55.75	63.62	65.15	78.89	73.54	72.63	75.02
Llama-2-13b-chat [0-shot]	36.92	58.18	73.78	57.01	56.47	42.31	38.03	51.75	21.08	38.29	64.10	78.19	73.46	71.92
Llama-2-13b-chat-finetune	66.11	68.64	**78.13**	67.45	70.08	62.34	66.48	60.02	47.75	59.15	76.62	71.88	68.44	72.31
-» Stanceformer	67.16	**71.43**	74.76	**73.98**	**71.83**	64.79	65.77	62.97	62.63	64.04	79.10	77.31	70.54	75.65

Results

Zero-shot dataset settings

- 입장별 성능 뿐만 아니라 전체적으로도 BERT 기반 모델 능가
- Fine-tuned 모델은 Zero-shot 추론 모델보다 |일반적으로 더 나은 성능
- Llama-2 모델은 모든 BERT 기반 모델보다 낮은 성능
- GPT-3.5은 Zero-shot 설정에서 기존의 최신 모델과Stanceformer를 능가

VAST-zero-shot	Pro	Con	Neu	All
BiCond†	44.6	47.4	34.9	42.8
CrossNet†	46.2	43.4	40.4	43.4
SEKT†	50.4	44.2	30.8	41.8
TPDG†	53.7	49.6	52.3	51.9
TOAD†	42.6	36.7	43.8	41.0
BERT†	54.6	58.4	85.3	66.1
TGA-Net†	55.4	58.5	85.8	66.6
BERT-GCN†	58.3	60.6	86.9	68.6
CKE-Net†	61.2	61.2	88.0	70.2
PT-HCL	56.1	62.8	87.9	68.9
-» Stanceformer	61.2	59.5	88.9	69.9
WS-BERT	57.0	**63.4**	**90.6**	70.3
-» Stanceformer	60.3	62.1	90.2	70.9
Closed-source LLM				
GPT-3.5 [0-shot]	**68.4**	63.2	81.9	**71.2**
Open-source LLM				
Llama-2-7b-chat [0-shot]	54.4	56.3	6.3	39.0
Llama-2-7b-finetune	49.5	34.0	46.5	43.4
-» Stanceformer	53.1	51.4	53.3	52.6
Llama-2-13b-chat [0-shot]	53.4	58.9	19.3	43.9
Llama-2-13b-finetune	49.0	27.7	48.2	41.6
-» Stanceformer	56.4	49.4	57.5	54.4

Analysis

- 다음의 세가지 방법을 비교
  - 기준 모델 (Targets Original)
  - 입력 '타겟' 정보가 모델로부터 완전히 숨겨짐 (Targets Masked)
  - Stanceformer를 적용한 모델
- 성능 하락이 작다는 것은 모델이 타겟들에 많은 주의를 기울이지 않는다는 것을 시사함

	SemEval-2016	
	BERT	BERTweet
Targets Original	61.31	64.23
Targets Masked	60.12	62.63
Stanceformer	**62.38**	**66.01**

Analysis

- SemEval-2016 데이터셋에 대해 학습된 BERTweet, GPT-3.5, Stanceformer가 생성한 샘플 예측

Text	Target	Ground Truth	BERTweet	GPT-3.5	Stanceformer
Remember, #God has it all worked out.	atheism	AGAINST	FAVOR	NONE	AGAINST
I am human. I look forward to the extinction of humanity with eager anticipation. We deserve nothing less.	atheism	AGAINST	NONE	NONE	AGAINST
I'm not a feminist, I believe in equality of the sexes! THATS EXACTLY WHAT FEMINISM IS	feminist movement	FAVOR	NONE	FAVOR	FAVOR
Men and women should have equal rights, we are all human...	feminist movement	FAVOR	NONE	FAVOR	FAVOR
Based on the long lines, I thought it was free burrito day at Pancheros but it was actually Hillary#ReadyForHillary	hillary clinton	FAVOR	AGAINST	AGAINST	FAVOR
Do you Progressives know how dangerously close you are to suppressing free speech? Stop it. #inners #readyforhillary	hillary clinton	FAVOR	AGAINST	AGAINST	FAVOR
I'm against abortion, gay marriage, AND Donald Trump for President. #gaymarriage #DonaldTrump	legalization of abortion	AGAINST	NONE	AGAINST	AGAINST
@toby_dorena Pregnant people have more than heartbeats. They have feelings, and the ability to make decisions about their health.	legalization of abortion	FAVOR	AGAINST	FAVOR	FAVOR
you can't say you support women's rights but be against abortion	legalization of abortion	AGAINST	AGAINST	FAVOR	FAVOR
Religions give its members an identity & without it, they cannot function. Feminists cannot function without feminism.	feminist movement	AGAINST	AGAINST	FAVOR	FAVOR

Table 5: Sample Predictions using BERTweet vs. GPT-3.5 vs. Stanceformer models for SemEval-2016 dataset. The text is highlighted as follows: CORRECT, INCORRECT predictions. Best viewed in color.

Generalization to Aspect-Based Sentiment Analysis domain

- 주어진 텍스트 내에서 특정 측면 용어들과 관련된 감성 극성(긍정, 부정 또는 중립)을 식별
- ABSA의 aspect들이 SD 작업의 target들과 유사
- BERTweet: REST16 측면에서 Macro-F1을 최대 4.1%, LAPT14 측면에서 정확도를 1.4% 향상
- BERT: LAPT14 측면에서 Macro-F1을 최대 2.2%, 정확도를 3.2% 향상
- Ablation Study
  - 타겟들이 마스킹되었을 때 2-4%의 성능 하락

Model	LAPT14	REST14	REST15	REST16
BERT				
Targets Original	62.56	**70.48**	59.22	60.71
Targets Masked	58.30	66.27	58.71	57.66
Stanceformer	**64.77**	70.11	**59.81**	**62.23**
BERTweet				
Targets Original	63.87	68.76	57.84	61.81
Targets Masked	59.37	66.42	57.68	61.02
Stanceformer	**67.00**	**71.97**	**58.46**	**65.88**

Model	LAPT14	REST14	REST15	REST16
BERT	78.76	90.09	84.26	88.84
-» Stanceformer	81.93	90.45	85.65	89.28
BERTweet	81.57	90.25	86.96	89.87
-» Stanceformer	**83.01**	**91.71**	**87.81**	**90.32**

Table 6: Accuracy scores across LAPT14, REST14, REST15, and REST16 datasets for BERT, BERTweet, and their Stanceformer variants.

Model	LAPT14	REST14	REST15	REST16
BERT	62.56	70.48	59.22	60.71
-» Stanceformer	64.77	70.11	**59.81**	62.23
BERTweet	63.87	68.76	57.84	61.81
-» Stanceformer	**67.00**	**71.97**	58.46	**65.88**

Table 7: Macro-F1 scores across LAPT14, REST14, REST15, and REST16 datasets for BERT, BERTweet, and their Stanceformer variants.

Challenges with LLMs

- LLM 파인튜닝은 상당한 컴퓨팅 자원을 필요로 함
- 모든 LLM을 파인튜닝할 수 있는 것은 아님
- LLM의 성능은 서로 다른 시드와 프롬프트에 따라 높은 변동성을 보이는 것으로 알려짐
- LLM 출력은 종종 통제가 불가능하여 관련 없거나 무의미한 텍스트를 생성
  - 특히 비하적이거나 유해한 텍스트 또는 논란이 되는 타겟들을 다룰 때 답변을 자제하는 경향
- LLM 평가가 완전히 투명하지 않을 수 있음
  - Stance Detection 작업의 테스트/훈련 세트에 대한 LLM 노출에 대해 우려
- LLM의 성능은 프롬프트의 품질에 크게 영향을 받음

Model	Prompt
GPT-3.5 [0-shot]	Following is a tweet. **<Tweet>**. Please predict the stance of the tweet towards target '**<Target>**'. Select from 'FAVOR', 'AGAINST' or 'NONE'.
Llama-2-7b-chat [0-shot] / Llama-2-13b-chat [0-shot]	<s>[INST] Following is a tweet. **<Tweet>**. Please predict the stance in the tweet towards the target **<Target>**. Answer in the form of pythonic dictionary. 'Stance': FAVOR/ AGAINST/NONE. Do not output anything else other than the dictionary. [/INST]
Llama-2-7b-chat [finetune] / Llama-2-13b-chat [finetune]	"<s>[INST] Consider the following input text. Please predict the stance in one word (FAVOR/ AGAINST/ NONE) in the input towards the target. Do not provide any justification. Input: **<Tweet>** Target: **<Target>** [/INST] Stance: **<Stance>**</s>"

Conclusion

- 기존의 transformer 모델들은 stance detection 작업에 포함된 타겟들을 효과적으로 우선순위화하는 능력이 부족함
- 이 연구에서는 향상된 타겟 인식을 특징으로 하는 stance detection에 특화된 transformer 아키텍처의 수정안인 Stanceformer를 제안
- BERT 기반 모델들과 자기회귀 LLM을 포함한 다양한 transformer 모델들을 사용하여 네 가지 stance detection 데이터셋에 대해 수행한 광범위한 실험
- 모든 설정에서 일관된 성능 향상을 보여줍니다.
- 우리는 Stanceformer가 감성분석과 같은 다른 도메인에도 잘 일반화됨을 보여줌

나의 생각

- 장점
  - Transformer를 target-specific하게 구조를 변경하는 방법을 제안하고 관련된 분야에 적용
  - LLM을 처음으로 stance detection을 위해 fine-tuning함
  - 그리고 그 결과를 분석하고 앞으로 활용에 있어 Challenge한 요소들을 정리함
  - 개인적으로 진행 중인 연구에서도 LLM이 RoBERTa 미세조정 성능을 넘지 못했는데, 비슷한 결과를 확인할 수 있어서 연구 진행에 도움이 됨
- 단점
  - Target-specific한 attention matrix가 특정 값을 target 부분에 일괄 더하는 것이라서 좀 단순하다고 생각
  - Tweet 데이터만 사용한 점

OPEN QUESTION

- Llama Fine-tuning이 BERT 기반 방법보다 안나온 이유?
- 제안한 방법 말고도 transformer 구조를 target-specific하게 만들 수 있는 방법이 있을까?
  - 아주대학교 형오 target-specific(토크나이저 단계)