EZ-STANCE: A Large Dataset for Zero-Shot Stance Detection

Chenye Zhao, Cornelia Caragea

Venue : ACL 2024

발제자 : 이다현 (hyundai@soongsil.ac.kr)

HUMANE Lab

2024-07-28

Background and Motivation

- **Stance Detection**
  - 자동으로 텍스트의 저자가 특정 Target에 대해 지지적인지, 비판적인지, 혹은 중립적인 입장인지 감지하는 것을 목표로 함
- 이전의 연구에서는 두가지의 Stance Detection task에 집중
  - In-target Stance Detection
    - 같은 target에 대한 데이터로 모델이 훈련되고 평가됨
  - Cross-target Stance Detection
    - 모델이 평가용 target과 관련이 있지만 구별되는 target으로 훈련된다
- 그러나, 모든 잠재적이거나 연관된 target을 훈련 세트에 통합하는 것은 비현실적임

Background and Motivation

Tweet	Nuclear Energy is a much safer and cost-efficient source of energy than coal and oil and people should be using it!
Stance/Noun-phrase targets	**Favor** / Nuclear Energy **Against** / Coal
Stance/Claim targets	**Favor** / Compared with traditional energy such as coal and gasoline, nuclear brings more security and is more economical. **Against** / Don't play with nuclear! We should stick with coal and fossil fuels. **Neutral** / Nuclear Energy will soon be the only energy left in the market. Coal and oil are outdated.

- "환경 보호" 도메인에 관련된 명사구 타겟과 주장 타겟의 예시

Background and Motivation

- **Zero-shot stance detection (ZSSD)**는 추론 단계에서 완전히 새로운 target에 대한 입장을 예측
- 현존하는 유일한 ZSSD 데이터셋인 **VAST**는 다음의 한계를 가짐
  1. 오직 명사구 target 만을 대상으로 하나, 실제로는 명사구 target과 주장 target 모두 나타남
  2. unseen target에 대한 입장을 예측하는 것을 목표로 하나,
     기존에는 unseen target들은 같은 domain에서 나온 target으로
     학습 단계의 target과 의미가 비슷해 덜 challenging함
  3. 중립 클래스에 대한 데이터를 기존 text와 target을 무작위로 변경함으로써 생성하여
     의미적 상관성이 결여됨

Contributions

- **EZ-STANCE**: a large **English Zero-shot stance detection** dataset collected from Twitter
  해당 데이터셋은 30606개의 주석 처리된 text-target 쌍으로 구성됨
- 두 개의 ZSSD subtask를 포함
  - Subtask A: target-based zero-shot stance detection
    - 모델은 unseen target + same domain으로 평가됨
  - Subtask B: domain-based zero-shot stance detection
    - 모델은 unseen target + different domain으로 평가됨
- 주석자들은 수작업으로 각 트윗에서 target을 추출하여 중립 클래스를 만듦
  - 트윗의 내용과 의미적 유사성을 보장
- 명사구 target과 주장 target을 모두 사용하여 보다 다양한 종류의 target을 포함
- 전통적인 모델과 사전학습 언어모델을 사용하여 EZ-STANCE가 challenging한 새로운 benchmark라는 것을 보임
- Stance Detection task를 NLI task로 변환하는 방법을 제안

Dataset Collection

- 2021년 5월 30일부터 2023년 1월 29일까지 트위터 API를 사용하여 50000개의 트윗을 수집
- 입장 검출에 적합하지 않은 키워드를 제거하기 위해 키워드 필터링을 수행
  - 주로 홍보를 목적으로 하는 프로모션 콘텐츠와 관련된 키워드를 제거
  - 사람들이 주로 단일 입장을 가지는 키워드를 필터링

YouTube shorts, modern history, work from home, herd immunity, living with covid, Fauci, public education, college football, pop culture, war, LGBTQ, environmental awareness, YouTube, career, vaccine, reels, democracy, pop culture, online shopping, hockey, reform, AI assistance writing, working class, election, parenting, global news, China, NBA, sports, student loan, traditional culture, Asian hate, presidential debate, Russia, bully, climate change, medicare, forcing electrical power, Mideast, doctors and patients, anti LGBTQ, post-covid, cooking, Snapchat, EU, presidential election, tictok, pfizer, business, general election, baseketball, prices, Chinese history, insurance, covid conspiracy, live shopping, SAT, Taliban, MLB, baseball, vaccine injury, tiger parents, environmental protection a, gency cultural output, Reels, government, family, new energy, WFH, clean energy, consumption concept, right wing, quality education, world news, stock market, private education, racism, long covid, NFL, vote, negative population growth, youtube, NASA, co-existence with Covid, WWE, DPR, political correctness, world cup, relationship, epidemic prevention, mideast, artificial intelligence, ethical consumption, Garbage classification, arming teachers, force kid to compete, health insurance, media, Negative population growth, terrorism, NATO, population aging, MLB's rule change, technology, wildfire, gun control, gender equality, migrant, doctors and patient, debate, mRNA vaccine, boxing, booster, leftists, republican, life in reels, abortion, teacher carry gun, Disney, overloaded kids, reward unreliable electricity gasoline price, international student, Ukraine, women's football, BLM, DPRK, privacy, shut down coal plants, homeschooling, physical education, men's football, NCAA, security, mask, sealed management, medical insurance, vegetarian, short video, iPhone, Iran, democrat, FDA, mid-term election, livestream shopping, CDC, women's rights, politic, electric vihicles, new york time, Hollywood, immigrant, Metoo, covid-19, equal rights, nuclear energy, mask mandate

Dataset Collection

- **8개 domain**의 논쟁적인 주제를 포함하는 **75개 키워드**를 선정
- “Covid 전염병" (CE), “세계 사건" (WE), “교육 및 문화" (EdC), “엔터테인먼트 및 소비" (EnC), “스포츠" (S), “권리" (R), “환경 보호" (EP), 그리고 “정치" (P)

Domain		Query Keywords
Covid Epidemic	CE	epidemic prevention, living with covid, herd-immunity, WFH, booster, vaccine, mask mandate, FDA, post-covid, Fauci
World Events	WE	world news, Ukraine, Russia, migrant, NATO, China, Mideast, negative population growth, terrorism
Education and Culture	EdC	public education, pop culture, cultural output, home schooling, AI assistance writing, arming teachers, private education, international student
Entertainment and Consumption	EnC	prices, gasoline price, online shopping, TikTok, iPhone, Reels, Disney, medical insurance, ethical consumption, vegetarian
Sports	S	World Cup, NBA, men’s football, women’s football, NCAA, MLB, NFL, WWE
Rights	R	gender equality, equal rights, women’s rights, LGBTQ, BLM, doctors and patients, racism, Asian hate, gun control
Environmental Protection	EP	climate change, clean energy, environmental awareness, environmental protection agency, shut down coal plants, nuclear energy, electric vehicle
Politics	P	government, republican, reform, leftists, democrat, democracy, right-wing, politic, presidential debate, presidential election, midterm election

Dataset Preprocessing

- 데이터셋 품질을 보장하기 위해 다음과 같은 전처리 단계를 수행
  - 단어 수가 20개 미만이거나 150개를 초과하는 트윗을 제거
  - 중복 트윗과 리트윗을 제거
  - 영어로 작성된 트윗만 남김
  - 광고 내용을 포함하는 트윗을 필터링
  - 이모지와 URL을 제거하여 노이즈를 줄임

- 각 키워드별로 약 86개의 트윗을 무작위로 선택하여 **총 6204개의 트윗**을 라벨링용으로 확보

Dataset Annotation

- 명사구 target에 대한 Annotation
  - 각 트윗에서 최소 2개의 명사구 target을 식별하도록 요구
  - 3명의 주석자에게 각 트윗-target 쌍에 입장 라벨을 할당하도록 지시
  - 6,204개의 트윗에 대해 **11,994개**의 트윗-명사구 target 쌍을 얻음

- 주장 target에 대한 Annotation
  - 트윗에 나타난 메시지를 기반으로 다음 세 가지 주장을 작성하도록 요구
    - **작성자가 주장 또는 메시지에 확실히 찬성하는 주장 (찬성)**
    - **작성자가 주장 또는 메시지에 확실히 반대하는 주장 (반대)**
    - **트윗의 정보만으로는 작성자가 주장 또는 메시지를 확실히 지지하거나 반대하는지 알 수 없는 주장 (중립)**
  - 더욱 도전적인 Task를 위해 다음과 같은 추가 요구사항을 설정
  - 찬성 주장은 트윗을 그대로 복제해서는 안되며, 반대 주장은 단순히 트윗 내용을 부정해서는 안 됨
  - 6,204개의 트윗에 대해 **18,612개**의 트윗-주장 target 쌍을 얻음

Example

	Noun-phrase targets			Claim targets		
Domain	Con	Pro	Neu	Con	Pro	Neu
CE	625	505	488	862	862	862
WE	557	367	540	772	772	772
EdC	395	538	436	731	731	731
EnC	429	601	703	945	945	945
S	125	516	500	625	625	625
R	574	660	340	786	786	786
EP	318	624	350	611	611	611
P	758	538	507	872	872	872
Overall	3,781	4,349	3,864	6,204	6,204	6,204

CE	Tweet	Cost of living off the scale, country being flooded with migrants, covid scam and jab injuries out there. How much more before the people decide enough is enough.
	N target/Stance	Covid Scams / Against
	C target/ Stance	Skyrocketing living costs and on the other side migrants will come in a lot of amounts so the country's population will increase someday. / Neutral
WE	Tweet	China's economy isn't just doing well. It is increasingly becoming 1 in several categories. Home prices are growing at slow and healthy rates, inflation is normal and healthy and the yuan is solid. The west should be trying to befriend China. Make a friend, not an adversary.
	N target/Stance	China's economy / Favor
	C target/ Stance	The economy of china is decreasing at an alarming rate due to which it's occupied last position in several categories. / Against
EdC	Tweet	To my Twitter pals who are parents in Ontario, trying to deal with homes chooling and work and all the stresses of the pandemic, my God, I don't know how you've managed to pull this off. But you have, even if you're exhausted. And you all rock.
	N target/Stance	home schooling / Against
	C target/ Stance	Parents in Ontario have managed to cope with homeschooling, work, and the pandemic, even if they are exhausted. / Favor
EnC	Tweet	Interviewer: why do you want this position? Me: so I can pay for all the online shopping I did this while being stressed about this interview.
	N target/Stance	online shopping / Favor
	C target/Stance	I do online shopping when I'm stressed. / Neutral
S	Tweet	Dwyane Wade winning an NBA Championship in his 3rd NBA season as the best player on the team .. does not get spoken on enough.
	N target/Stance	Dwyane Wade / Favor
	C target/Stance	Dwyane Wade's success in his 3rd NBA season made him the best player of all times. / Neutral
R	Tweet	The FEUHS Student Government is one with the LGBTQIA community in celebrating the PrideMonth2021 and pursuing equal rights for everyone, regardless of sexual orientation, gender identity, and expression.
	N target/Stance	Equal Rights / Favor
	C target/ Stance	Regardless of sexual orientation, gender identity, or gender expression, the FEUHS Student Government opposes equitable rights for everyone. / Against
EP	Tweet	The Sines coal plant in Portugal has been shut down nine years ahead of schedule, reducing the country s carbon emissions by 12%. A second and final plant is due to close in November which will make Portugal the fourth European country to eliminate.
	N target/Stance	Carbon emissions / Against
	C target/ Stance	Portugal's Sines coal facility was shut down nine years earlier than expected, cutting the nation's carbon emissions by 12 percent. / Favor
P	Tweet	I wish Democrats would play tough and just release an ad that says "GOP loves guns more than our kids." Just show the 234 mass shootings in 2022 and how GOP has obstructed every attempt at gun reform. There's no lie in that claim. At the very least don't call them "rational."
	N target/Stance	GOP / Against
	C target/Stance	The GOP will bring gun reform to stop the mass shootings. / Neutral

Dataset Split

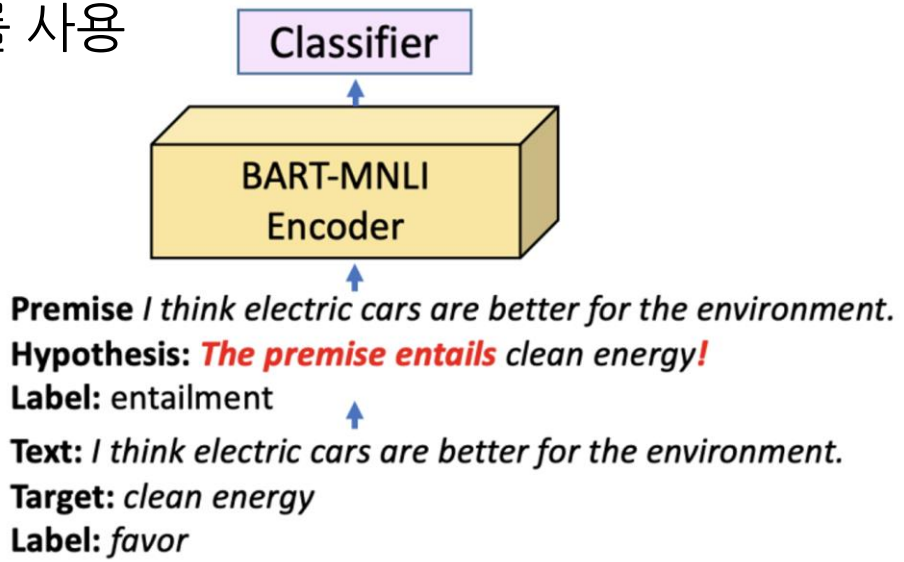
- Subtask A와 Subtask B 모두 zero-shot 세팅을 보장하기 위해
  training, validation, test set은 트윗과 target을 공유하지 않음

- Subtask B에서는 7개의 domain 데이터(source)를 training과 validation에 사용하고,
  나머지 한 개의 도메인(zero-shot)을 test에 사용

		# Examples		# Unique			Avg. Length			Lexsim
		N	C	N	C	T	N	C	T	(%)
Subtask A	Train	8,705	12,264	4,842	12,248	4,088	1.8	18.4	39.8	-
	Val	1,667	3,081	1,578	3,078	1,027	2.3	19.2	39.1	13
	Test	1,622	3,267	1,613	3,253	1,089	2.3	18.9	39.3	12
Subtask B (Covid Epidemic)	Train	8,498	13,167	5,875	13,151	4,389	2	18.6	39.3	-
	Val	1,231	2,754	1,220	2,744	918	2.3	18.4	39.5	11
	Test	1,716	2,607	1,156	2,602	869	1.9	18.7	41.1	10

Methodology

- 문서와 타겟을 자연어 추론 태스크에서 전제와 가설로 바꾸는 방법을 제안
- 특히, 명사구 타겟을 보다 정교한 가설로 공식화하기 위해 두개의 프롬프트 템플릿을 이용
  - *"The premise entails [target]!"*
  - *"The premise entails the hypothesis [target]!"*
- 각 명사구 타겟에 대해 우리는 무작위로 둘 중 하나를 적용
- 주장 타겟의 경우, 이미 가설의 형태와 매우 유사하기 때문에 따로 프롬프트를 적용하지 않음
- MNLI 데이터셋에 사전학습된 BART-large Encoder를 사용



Baselines and Models

- BiCE (Augenstein et al., 2016)
- CrossNet (Xu et al., 2018)
- TGA-Net (Allaway and McKeown, 2020)

- Transformer 기반 모델의 기본 버전을 미세 조정
  - BERT (Devlin et al., 2019)
  - RoBERTa (Liu et al., 2019)
  - XLNet (Yang et al., 2019)
- NLI 사전 학습 모델을 평가하기 위해, 다음 방법들을 비교
  - BART-MNLI-ep: 명사구 대상에 적용된 제안 프롬프트를 사용하여 EZ-STANCE 데이터셋을 이용
  - BART-MNLI-e: 프롬프트 없이 원래의 EZ-STANCE 데이터셋을 이용
  - BART-MNLI: 사전 학습된 BART-MNLI 모델을 인코더와 디코더 모두 사용, 미세 조정X

Experiments

1. Subtask A와 Subtask B에 대한 실험을 수행
2. EZ-STANCE와 VAST 데이터셋을 비교
3. 다양한 프롬프트의 영향 연구
4. 명사구 target과 주장 target 을 하나의 데이터셋으로 통합한 효과 탐구
5. 주장 target 에 대한 편향 분석을 수행
    - 클래스별 F1 점수와 모든 클래스에 걸친 매크로 평균 F1 점수를 평가 메트릭으로 사용

Experiments

- **Subtask A: Target-based Zero-shot Stance Detection**

목표는 완전히 보지 못한 target 을 기반으로 분류기를 평가하는 것

혼합된 대상(명사구와 주장 모두)을 포함한 전체 데이터셋, 명사구 대상만 포함한 데이터셋, 주장 대상만 포함한 데이터셋을 각각 사용하여 수행

	Mixed targets				Noun-phrase targets				Claim targets			
	Con	Pro	Neu	All	Con	Pro	Neu	All	Con	Pro	Neu	All
BiCE	.539	.358	.536	.478	.583	.550	.453	.529	.313	.346	.317	.325
Cross-Net	.504	.485	.571	.520	.559	.552	.466	.526	.473	.448	.622	.514
TGA Net	.558	.564	.625	.582	.641	.603	.503	.583	.514	.551	.687	.584
BERT	.724	.732	.756	.738	.669	.619	.535	.608	.706	.768	.872	.782
RoBERTa	.787	.785	.769	.780	.712	.677	.529	.639	.821	.856	.881	.853
XLNet	.767	.766	.760	.764	.685	.652	.531	.623	.806	.841	.880	.842
BART-MNLI	.652	.699	.632	.661	.194	.531	.205	.310	.789	.832	.783	.801
BART-MNLI-e	.816	.808	.773	.799	.729	.690	.542	.653	**.858***	**.888***	**.892***	**.879***
BART-MNLI-c$_p$	**.818***	**.813***	**.783***	**.805***	**.739***	**.692***	**.576***	**.669***	-	-	-	-

Experiments

- Subtask B: Domain-based Zero-Shot Stance Detection
  - 완전히 새로운 도메인에서 보지 못한 주제를 사용하여 분류기를 평가하는 것에 중점을 둠
  - 특히, 하나의 도메인을 제로샷 도메인으로 선택하고 나머지 7개 도메인을 출처 도메인으로 사용
  - 출처 도메인의 데이터를 사용하여 모델을 훈련 및 검증하고, 제로샷 도메인의 데이터를 사용하여 모델을 테스트

Model		CE	WE	EdC	EnC	S	R	EP	P
BiCE	M	.441	.443	.480	.451	.458	.485	.465	.439
	N	.461	.485	.486	.476	.434	.515	.514	.433
	C	.323	.313	.325	.319	.324	.309	.319	.310
CrossNet	M	.482	.489	.501	.484	.470	.531	.489	.484
	N	.471	.502	.489	.487	.487	.505	.522	.476
	C	.495	.495	.499	.486	.475	.505	.473	.501
TGA-Net	M	.535	.545	.565	.559	.553	.606	.570	.562
	N	.471	.528	.552	.544	.530	.565	.558	.552
	C	.572	.568	.595	.591	.545	.610	.567	.567
BERT	M	.681	.689	.716	.685	.698	.728	.695	.698
	N	.567	.560	.580	.577	.587	.612	.578	.569
	C	.753	.760	.784	.763	.769	.780	.764	.765
RoBERTa	M	.716	.728	.759	.744	.738	.763	.736	.746
	N	.612	.600	.629	.596	.598	**.633**	.625	.591
	C	.815	.833	.856	.845	.833	**.831**	.825	.828
XLNet	M	.707	.722	.741	.724	.719	.745	.734	.717
	N	.586	.609	.596	.588	.581	.622	.605	.580
	C	.790	.796	.832	.829	.793	.819	.808	.802
BART-MNLI	M	.590	.591	.633	.627	.656	.616	.638	.577
	N	.314	.270	.336	.334	.368	.330	.377	.309
	C	.755	.797	.794	.787	.788	.780	.768	.752
BART-MNLI-e	M	.751	.758	.771	.769	.766	.765	.759	.757
	N	.604	**.620**	**.639**	.609	.582	.624	.623	.610
	C	**.850***	**.866***	**.874***	**.866***	**.866***	.830	**.850***	**.846***
BART-MNLI-e$_p$	M	**.752***	**.769***	**.772***	**.771***	**.768***	**.783***	**.768***	**.763***
	N	**.613**	.613	.629	**.613***	**.613***	.628	**.638***	**.613***
	C	-	-	-	-	-	-	-	-

EZ-STANCE vs VAST

- **대상 다양성**
  - 훨씬 더 많은 제로샷 대상을 포함
  - 더 다양한 제로샷 대상에 일반화될 수 있음을 시사
- **과제의 난이도**
  - 하나의 데이터셋으로 모델을 훈련시키고 다른 데이터셋으로 테스트하는 교차 데이터셋 실험을 수행
  - 모델이 인-데이터셋 설정에서 교차 데이터셋 설정보다 훨씬 높은 성능을 보임
  - VAST로 훈련된 모델이 EZ-STANCE 테스트 세트의 중립 클래스에서 매우 저조한 성능을 보이는 반면, EZ-STANCE로 훈련된 모델은 VAST에서 훨씬 높은 성능을 보입

	Subtask A			Subtask B (CE)		
	Train	Val	Test	Train	Val	Test
V	4,003	383	600	-	-	-
E	17,090	4,656	4,866	19,026	3,964	3,758

Train/Val	Test	Con	Pro	Neu	All
E	V	.578	.626	.286	.497
V	E	.644	.615	.005	.421
E	E	.739	.692	.576	.669
V	V	.719	.701	.919	.780

Impact of Prompt Templates

- 선택한 2개의 프롬프트가 다른 프롬프트보다 더 우수한 성능을 보임
- 두 가지 프롬프트 중 하나를 무작위로 선택한 접근 방식이 한 가지 유형의 프롬프트만을 사용한 모델보다 더 나은 성능

Prompts	Con	Pro	Neu	All
The premise entails [target]!	.729	.684	**.578**	.664
The premise entails the hypothesis [target]!	.727	.688	.571	.662
I am in favor of [target]!	.727	.690	.559	.659
I support [target]!	.728	.690	.551	.656
I am against [target]!	.718	.678	.558	.652
I disagree with [target]!	.721	.686	.567	.658
Ours	**.739**	**.692**	.576	**.669**

Impact of Incorporating Two Target Types

- 명사구 target으로 훈련된 모델을 주장 target으로 평가하고, 그 반대로도 실험
  - 명사구 대상과 주장 대상을 하나의 데이터셋에 통합할 필요성을 탐구하기 위함

Train/Val	Test	RoBERTa	BART-MNLI-e
M	N	.609	.619
M	C	.859	.880
C	N	.364	.349
N	C	.309	.325

Spuriosity Analysis for Claim Targets

- 주장 target을 기반으로 한 입장을 단순히 주장에만 의존하여 감지할 수 없는지 확인하기 위해 편향 분석을 수행

Train/Val	Test	RoBERTa	BART-MNLI-e
M	N	.609	.619
M	C	.859	.880
C	N	.364	.349
N	C	.309	.325

Conclusion

- 대규모 영어 ZSSD 데이터셋인 EZ-STANCE를 소개
- 기존 유일한 ZSSD 데이터셋인 VAST와 비교하여, 더 크고 더 도전적임을 확인
- EZ-STANCE는 명사구 target 과 주장 target을 모두 포함하며,
  두 가지 도전적인 ZSSD 하위 과제인 target 기반 ZSSD와 도메인 기반 ZSSD를 포괄
- 텍스트에서 대상을 추출하여 중립 클래스의 데이터 품질을 향상
- ZSSD baseline을 통해 EZ-STANCE를 평가하고, ZSSD를 NLI 작업으로 변환하여 전통적인 baseline보다 우수한 성능을 보임을 제안

나의 생각

- **장점**
  - Target의 유형에 따라 Stance 분류 성능이 달라진다는 점을 포착
  - NLI task와 Stance Detection task 사이의 연관성 입증
- **단점**
  - 여전히 주석 처리된 데이터셋에 의존하는 부분이 있음
    - 전혀 다른 도메인에 대한 평가 성능이 제한적
  - 데이터셋이 주로 소셜 미디어에서 수집되어 다른 유형의 텍스트에는 유효하지 않을 듯 함
    - 뉴스, 에세이 등

Open question

- 문장 단위의 Stance Detection을 넘어 문서의 편향을 탐지할 수 있을까?
  할 수 있다면 라벨링된 데이터셋, 적합한 모델 개발 등 다양한 요소 중
  어떤 것이 성능에 가장 긍정적인 영향을 미칠까?
- 학습된 텍스트의 특성을 넘어선 범용적인 Stance Detection을 위해서는
  Task를 어떻게 구성해야 할까?