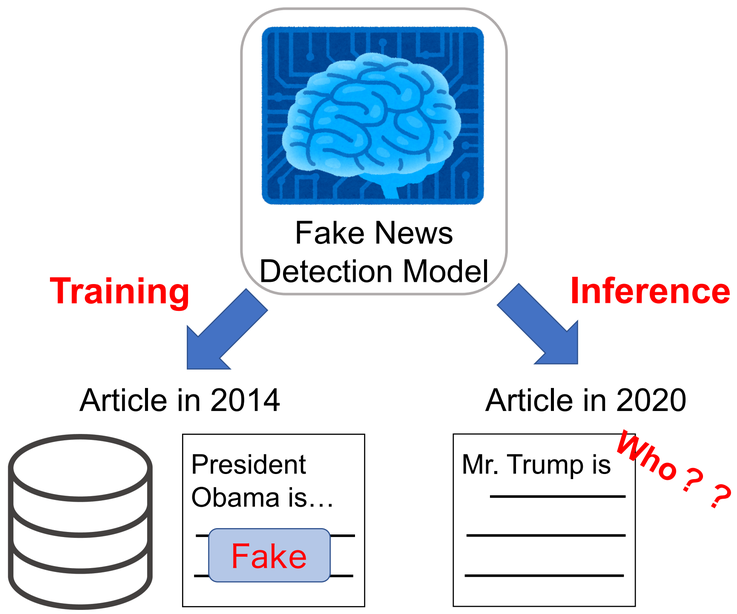
Mitigation of Diachronic Bias in Fake News Detection Dataset

Taichi Murayama, Shoko Wakamiya, Eiji Aramaki
(Nara Institute of Science and Technology)

Overview



- We call '**Diachronic Bias**', which is a possibility to make incorrect judgments regardless of the veracity of the input sentence, when the lexical information of input articles differs greatly from the text the model was trained on.
- Examining word and label deviation in fake news datasets.
- Validate **the proposed masking methods using Wikidata** to mitigate bias.

Background

- Increasing the research on building fake news detection models
- Most datasets for training the model consist of fake news diffusion in the real world.
- Fake news on different topics spread at different times.
  - 2013：President Obama
  - 2016：the US 2016 election
  - 2020：COVID-19
- A model trained on the corpus from a specific period has the possibility to make incorrect detection of the veracity of the input when the input includes a new word or knowledge.

Investigate **multiple masking methods to mitigate diachronic bias** and build a robust fake news detection model for out-of-domain data.

Dataset and the Deviation

Datasets

Four fake news detection datasets with two labels, real and fake.

- MultiFC[1] : Consists of news before 2015
- Horne17[2] : Consists of news in the US 2016 election
- Celebrity[3] : Consists of celebrity news in 2016, 2017
- Constraint[4] : Consists of COVID-19 news

Deviation of phrase

Investigating the correlation between phrase appearance and labels to examine bias in each fake news dataset
⇒ Using **Local Mutual Information (LMI)**

$$LMI(w,l) = p(w,l) \cdot \log\left(\frac{p(l|w)}{p(l)}\right)$$

w: phrase, l: label, $p(l|w)$: conditional probability
$p(w,l)$: joint distribution between phrase and label

Top 10 phrases of the highest correlation

MultiFC -2015

Real			Fake		
Bigram	**LMI**	$p(l\|w)$	**Bigram**	**LMI**	$p(l\|w)$
mitt romney	218	0.69	health care	631	0.64
if you	217	0.70	**barack obama**	365	0.69
rhode island	190	0.75	**president barack**	337	0.70
new jersey	177	0.67	**scott walker**	258	0.81
john mccain	167	0.73	says president	218	0.78
no. 1	128	0.86	care law	185	0.80
voted against	128	0.71	will be	162	0.63
any other	125	0.61	**hillary clinton**	159	0.67
does not	119	0.71	**gov. scott**	148	0.72
this year	116	0.75	social security	144	0.68

Horne17 2016

Real			Fake		
Bigram	**LMI**	$p(l\|w)$	**Bigram**	**LMI**	$p(l\|w)$
trump has	112	0.82	**donald trump**	605	0.42
national security	106	0.88	**hillary clinton**	440	0.50
would be	104	0.72	i think	292	0.68
people who	92	0.89	united states	258	0.51
transition team	88	1.0	have been	230	0.41
mr. trump	80	0.94	**bill clinton**	208	0.70
smug style	77	1.0	we are	206	0.56
george w.	76	0.90	**hillary clinton's**	187	0.58
republican party	76	0.91	**president obama**	171	0.55
new york	70	0.77	**ted cruz**	149	0.80

- In MultiFC, ex-president '**barack obama**' has high correlation with fake label, while in Horne17, '**hillary clinton**' and '**donald trump**' have high correlation.
- In particular, real labels tend to be highly correlated with **general terms**, while fake labels tend to be highly correlated with **person names**.

Examination of Mitigation methods

Mitigation methods

Investigating multiple masking methods to mitigate diachronic bias and build robust models for out-of-domain data.

- NE Deletion: Remove the word tagged as Named Entity (NE)
- Basic NER: Replace the tagged word NE tag name
- WikiD: Replace the word tagged as PER tag, one of NE tags, public position or occupation tag id in Wikidata
  e.g. Obama = Trump = Q11696
- WikiD+Del: After applying WikiD rule, remove other tagged words
- WikiD+NER: After applying WikiD rule, replace other tagged words NE tag name

Lexicalized	18 states including **US UK** and **Australia** request PM **Modi** to head a task force to stop coronavirus
NE Deletion	18 states including and request PM to head a task force to stop coronavirus
Basic NER	18 states including **LOC LOC** and **LOC** request PM **PER** to head a task force to stop coronavirus
WikiD	18 states including **US UK** and **Australia** request PM **Q22337580** to head a task force to stop coronavirus
WikiD+Del	18 states including and request PM **Q22337580** to head a task force to stop coronavirus
WikiD+NER	18 states including **LOC LOC** and **LOC** request PM **Q22337580** to head a task force to stop coronavirus

Experimental setting

Model: BERT_base
Data: train 80%, test 20%
Evaluation: Accuracy

In-domain results

- No mask achieves the highest score in almost dataset
- There is only a few points difference compared to other masking methods

	MultiFC	Horne17	Celebrity	Constraint
No Mask	**0.681**	**0.746**	**0.760**	0.960
NE Del	0.656	0.706	0.750	0.959
Basic NER	0.659	0.735	0.750	0.950
WikiD	0.675	0.725	0.730	**0.967**
WikiD+Del	0.660	0.706	0.700	0.959
WikiD+NER	0.660	0.640	0.730	0.957

Out-domain results

- Most masking methods achieve higher score in many out-of-domain dataset, than No mask.
- **NE Deletion and Basic NER** perform higher accuracy in **9 (out of 12)** experimental settings and **WikiD and WikiD+Del** perform higher accuracy in **10 (out of 12)** experimental settings, than No mask.

E.g.

		MultiFC	Horne17	Celebrity	Constraint
MultiFC	No Mask	-	0.706	**0.660**	0.530
	NE Del	-	0.706	0.590	*0.664
	Basic NER	-	0.725	0.600	*0.680
	WikiD	-	**0.746**	0.590	***0.689**
	WikiD+Del	-	0.725	**0.660**	*0.669
	WikiD+NER	-	0.632	0.520	*0.667

[1] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In Proc. of EMNLP-IJCNLP, pp. 4677–4691, 2019.
[2] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proc. of ICWSM, Vol. 11, 2017.
[3] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fakenews. InProc. of COLING, pp. 3391–3401, 2018.
[4] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Ami-tava Das, Fighting an infodemic:Covid-19 fake news dataset.arXiv:2011.03327, 2020