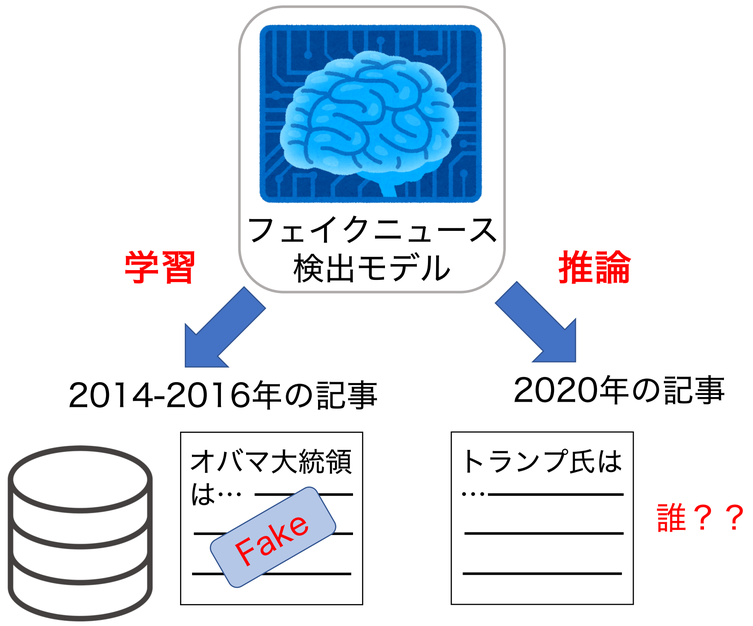
P5-17

フェイクニュース検出データセットにおける通時的バイアス

村山太一，若宮翔子，荒牧英治
(奈良先端科学技術大学院大学)

Overview



- モデルが学習したテキストと語彙情報が大きく異なる入力がされると，その文章の真偽に関わらず判定が誤る可能性を「通時的バイアス」と呼称
- 多く利用されるフェイクニュース検出データセットで語彙とラベルの偏りを確認
- Wikidataを活用した，偏りを緩和するマスキング手法を提案し有効性を確認

背景

- フェイクニュース検出モデル構築の研究が増加
- 検出モデルの学習に用いられるデータセットは，現実のフェイクニュースから構成
- 現実のフェイクニュースは時期によって異なるフェイクニュースが流行
  - 2013年：オバマ大統領
  - 2016年：米国大統領選挙
  - 2020年：COVID-19
- 特定の時期のテキストによって学習したモデルは新しい語彙が入力されるとその文章の真偽の判定を誤る可能性
- データセットの作成時期に依存することから「通時的バイアス」と呼称

通時的バイアスを緩和し，ドメイン外のデータに対しても頑健なフェイニュース検出モデル構築のため複数のマスキング手法を検討

データセットとその偏り

データセット

FakeとRealの2つのラベルを持つ4つのフェイクニュース検出データセットを利用

- MultiFC[1]：2015年以前のニュースで構成
- Horne17[2] ：2016年米大統領選挙のニュースで構成
- Celebrity[3]：2016年，2017年の芸能ニュースで構成
- Constraint[4] ：COVID-19関連のニュースで構成

データセットの偏り

各データセットの偏りを検証するため，データセットごとの語彙とラベルの偏りを検証
⇒ Local Mutual Informationを利用して検証

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

w: フレーズ, l: ラベル
$p(w,l)$: フレーズとラベルの同時確率，$p(l|w)$: 条件付き確率

各ラベルとの相関が高い上位10フレーズ

MultiFC ～2015年

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年

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

- MultiFCでは“barack obama”といった当時の大統領がFakeラベルとの相関が高い．一方で，2016年大統領選挙の記事のHorne17では大統領候補であった“hillary clinton”や“donald trump”がFakeラベルと高い相関
- 特にRealラベルは一般的な用語と，Fakeラベルは人名と高い相関を持つ傾向

通時的バイアス緩和の検討手法

緩和手法

通時的バイアスを緩和しドメイン外のデータにも頑健なモデル構築のために，複数のマスキング手法を検討

- NE Deletion: Named Entity (NE)とタグ付けされた語彙を削除
- Basic NER: NEのタグに語彙を置き換え
- WikiD: NEタグがPERのものをWikidataの公的な地位 or 職業のタグに置き換え,
  i.e. Obama = Trump = Q11696
- WikiD+Del: WikiDのルールに加え，他のNEを削除
- WikiD+NER: WikiDのルールに加え，他のNEをタグに置き換え

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

実験設定

モデル: BERT_base
データ: train 80%, test 20%
評価: Accuracy

In-domainでの結果

- マスキング手法を用いたい手法が最も高い精度
- しかし，他のマスキング手法と比較して数ポイントの差しか存在しない

	MultiFC	Horne17	Celebrity	Constraint
Lexicalized	**0.681**	**0.746**	**0.760**	0.960
NE Deletion	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での結果

- そのままの入力よりもマスキング手法の方が多くのドメイン外データで高い精度を達成
- すべてで12のドメイン外データでの検証を行った．そのうち，NE Deletion/Basic NERでは9/12, WikiD/WikiD+Delでは10/12の実験設定で通常の入力より高い精度

例

データ		MultiFC	Horne17	Celebrity	Constraint
MultiFC	Lexicalized	-	0.706	**0.660**	0.530
	NE Deletion	-	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