Progressive Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis

Jialong Tang, Ziyao Lu, Jinsong Su, Yubin Ge, Linfeng Song, Le Sun, Jiebo Luo

発表者：豊田工業大学　D2　辻村有輝

Model	Sentence	Ans./Pred.
TNet-ATT	The [folding chair] i was seated at was uncomfortable .	Neg / Neu
TNet-ATT(+AS)	The [folding chair] i was seated at was uncomfortable .	Neg / Neg
TNet-ATT	The [food] did take a few extra minutes ... the cute waiters ...	Neu / Pos
TNet-ATT(+AS)	The [food] did take a few extra minutes ... the cute waiters ...	Neu / Neu

- Aspect-levelの極性分類
- 訓練中に得られるアテンションの注目量からアテンション教師を自動作成し再学習
- 既存のSotAなモデルに導入し性能向上

スライド中の図・表は論文内から引用

背景

- 最近のAspect-levelの極性分類ではアテンションベースのNNモデルが支配的
- ただしアテンションは高頻度単語に過度に注目してしまいやすい問題がある
- 例：訓練中にはsmallが出現するnegativeな事例が多い➡smallに注目が強く集まる

Type	Sentence	Ans./Pred.
Train	The **[place]** is small and crowded but the service is quick .	Neg / —
Train	The **[place]** is a bit too small for live music .	Neg / —
Train	The service is decent even when this small **[place]** is packed .	Neg / —
Test	At lunch time , the **[place]** is crowded .	Neg / Pos
Test	A small area makes for quiet **[place]** to study alone .	Pos / Neg

  - 最初の例のcrowdedも同じくnegativeの根拠となる単語のはず
    ➡ test中のcrowdedが出現する文で注目できず間違える
  - 逆にpositiveの文でsmallに注目しすぎて間違える

背景

- アテンションは高頻度単語に過度に注目してしまいやすい問題がある

  ➡ アテンションに対する教師を用意することで性能向上に繋げられる可能性がある

- アテンションの教師を用意する既存研究
  - Neural machine translation with supervised attention (Liu et al, 2016)
  - Exploiting argument information to improve event detection via supervised attention mechanisms (Liu et al, 2017)
  - Who is killed by police: Introducing supervised attention for hierarchical lstms (Nguyen and Nguyen, 2018)
  - これらで用意されるアテンションの教師は人手で作られておりコストが高い

➡ モデルの学習したアテンションの注目量から自動的に教師を作成する

アイディア

アテンションに対する教師を用意することで性能向上に繋げられる可能性がある

- モデルが最大の注目を当てた単語が出力に最大の影響を与える（はず）
  - 正解時に最も注目した単語はactiveな単語として注目を当て続けていくべき
  - 不正解時に最も注目を集めた単語はmisleadingな結果を導く単語として注目を外すべき

➡ 最も注目した単語を記録しておくことでアテンションの教師を作成できる

- モデルは出現頻度の高い単語に注目を集めがち
  - 根拠となりうる低頻度語がアテンションから隠されてしまう（例：crowded）

➡ 最も注目した単語を一旦隠して再学習することで低頻度語にも注目が当たるようになる

➡ 「最も注目した単語を記録する→それらをマスクして再学習する」ことを繰り返しactive/misleadingな単語を集め，アテンションの教師として利用する

提案手法｜アルゴリズム

1. 通常の学習を行う
2. 以下をK回繰り返す
    1. 各訓練事例xに対しアテンション$\alpha(x)$を計算し，そのエントロピー$E\big(\alpha(x)\big)$を計算
    2. アテンションのエントロピーが閾値ϵ_αを下回ったとき
        1. 正解の場合は最注目した単語をその入力におけるactiveな単語集合S_aに登録
        2. 不正解の場合はその入力におけるmisleadingな単語集合S_mに登録
        3. 登録されたactive/misleadingな単語は<mask>トークンで置き換える
            （登録・<mask>への置き換えは事例ごとに独立で，他の事例に影響しない）
    3. マスクされた入力を使い再学習を行う（パラメータは以前の学習後から引き継ぐ）
3. 元々の訓練事例でアテンションに教師をつけ学習（2ページ後で説明）

イテレーションごとの入力・アテンション例

Iter	Sentence	Ans./Pred.	$E(\alpha(x'))$	x'_m
1	The **[place]** is small and crowded but the service is quick .	**Neg / Neg**	**2.38**	***small***
2	The **[place]** is ⟨mask⟩ and crowded but the service is quick .	**Neg / Neg**	**2.59**	***crowded***
3	The **[place]** is ⟨mask⟩ and ⟨mask⟩ but the service is quick .	**Neg / Pos**	**2.66**	***quick***
4	The **[place]** is ⟨mask⟩ and ⟨mask⟩ but the service is ⟨mask⟩ .	**Neg / Neg**	**3.07**	—

（エントロピーの閾値を3に設定）

1. smallが最注目となり予測も正解➡smallをactiveとしてマスク
2. crowdedが最注目となり予測も正解➡crowdedをactiveとしてマスク
3. quickが最注目となり予測も正解➡quickをmisleadingとしてマスク
4. アテンションのエントロピーが閾値を上回ったため根拠単語は抽出されない

提案手法｜アテンションに対する教師

元々の訓練事例でアテンションに教師をつけ学習

- 入力：マスクされていない訓練事例
- モデルのアテンションを矯正する正則化項を目的関数に導入して学習

$$J_s(D_s;\theta) = -\sum_{(x,t,y)\in D_s} \{\underbrace{J(x,t,y;\theta)}_{\text{通常のタスク損失}} + \gamma \underbrace{\triangle(\alpha(s_a(x)\cup s_m(x)),}_{\text{出力アテンション}} \underbrace{\hat{\alpha}(s_a(x)\cup s_m(x));\theta)}_{\text{教師アテンション}}\}$$

  - active/misleadingな単語に基づいた教師アテンションと出力アテンションのユークリッド距離を近づける
    - active： $\frac{1}{|s_a(x)|}$　　misleading：0

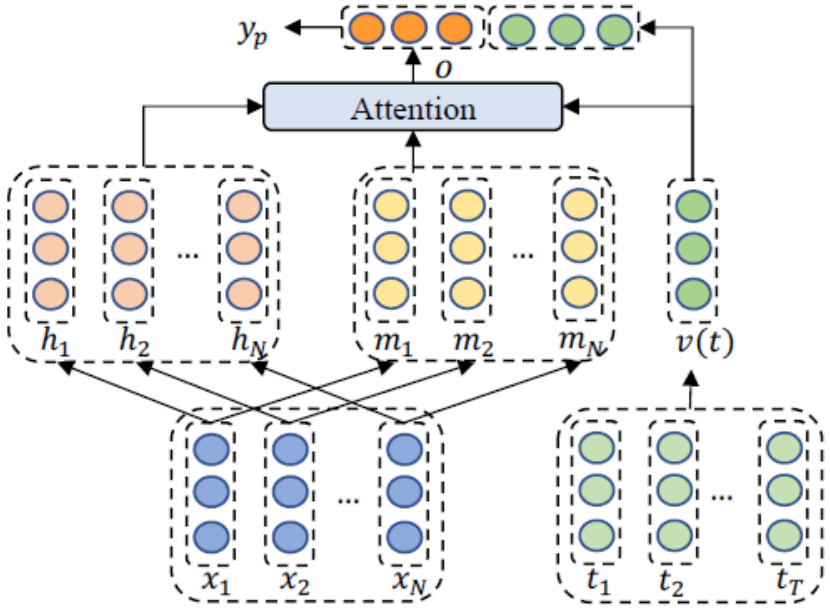
入力	The [place] is small and crowded but the service is quick .	赤：active
教師アテンション $\hat{\alpha}$	0.5 0.5 0	青：misleading

- 以前の学習済みパラメータは引き継がず，再初期化して学習する

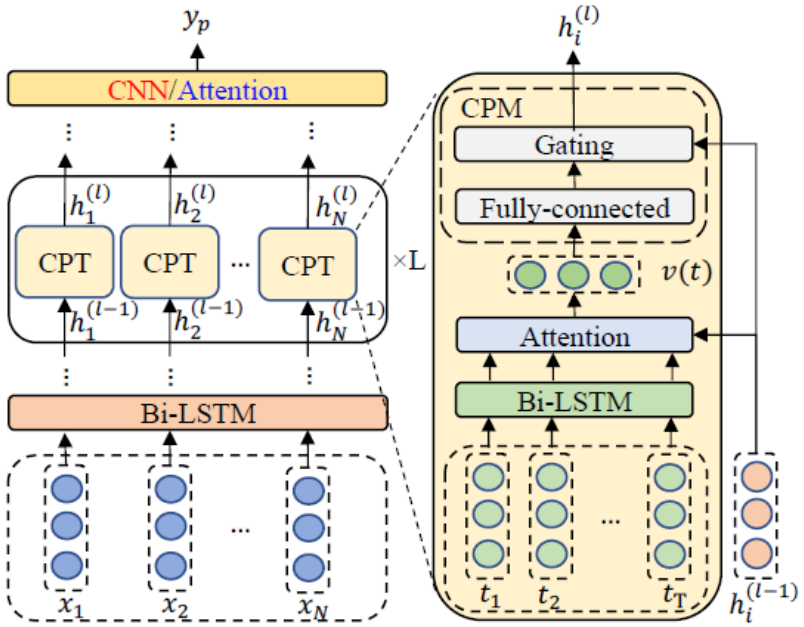
実験

既存のSotAなモデルのアテンション部分に提案手法を導入する

- TNetに関しては最終層のCNNをAttentionに変更（TNet-ATT）



Memory Network (MN)

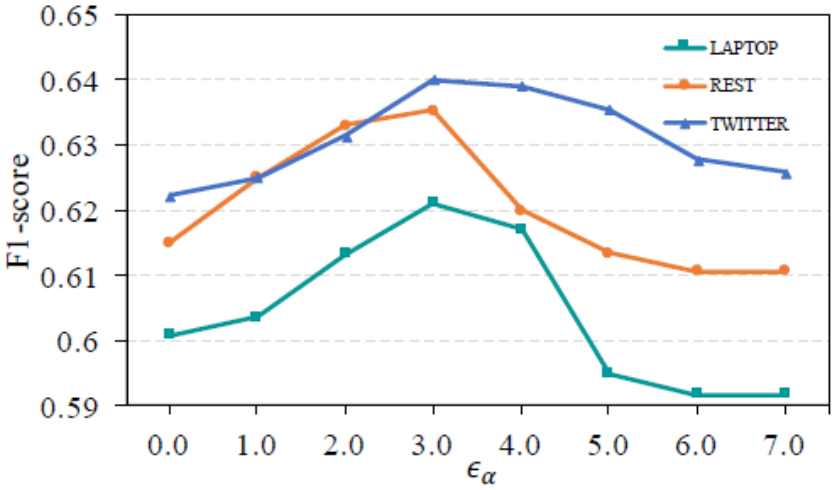


Transformation Network (TNet)

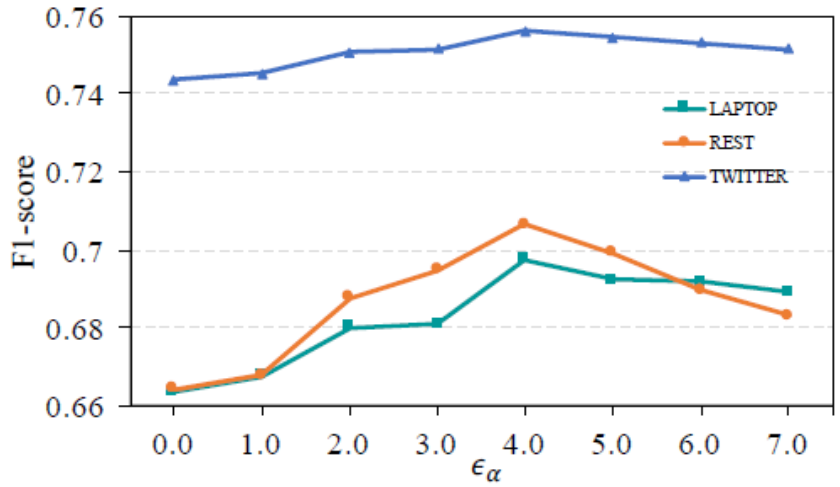
実験結果｜スコア

Model	LAPTOP		REST		TWITTER	
	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy
MN (Wang et al., 2018)	62.89	68.90	64.34	75.30	—	—
MN	63.28	68.97	65.88	77.32	66.17	67.71
MN(+KT)	63.31	68.95	65.86	77.33	66.18	67.78
MN(+AS_m)	64.37	69.69	68.40	78.13	67.20	68.90
MN(+AS_a)	64.61	69.95	68.59	78.23	67.47	69.17
MN(+AS)	**65.24****	**70.53****	**69.15****	**78.75***	**67.88****	**69.64****
TNet (Li et al., 2018)	71.75	76.54	71.27	80.69	73.60	74.97
TNet	71.82	76.12	71.70	80.35	76.82	77.60
TNet(+KT)	71.74	76.44	71.36	80.59	76.78	77.54
TNet-ATT	71.21	76.06	71.15	80.32	76.53	77.46
TNet-ATT(+KT)	71.44	76.06	71.01	80.50	76.58	77.46
TNet-ATT(+AS_m)	72.39	76.89	72.04	80.96	77.42	78.08
TNet-ATT(+AS_a)	73.30	77.34	72.67	81.33	77.63	78.47
TNet-ATT(+AS)	**73.84****	**77.62****	**72.90****	**81.53***	**77.72****	**78.61***

青：ベースライン　　赤：提案手法



MNにおけるエントロピー閾値の効果



TNetにおける閾値の効果

実験結果｜学習後のアテンション

Model	Sentence	Ans./Pred.
TNet-ATT	The [folding chair] i was seated at was uncomfortable .	Neg / Neu
TNet-ATT(+AS)	The [folding chair] i was seated at was uncomfortable .	Neg / Neg
TNet-ATT	The [food] did take a few extra minutes ... the cute waiters ...	Neu / Pos
TNet-ATT(+AS)	The [food] did take a few extra minutes ... the cute waiters ...	Neu / Neu

- １番目の例ではUncomfortableの出現頻度の低いせいで注目が集まらず，ベースラインモデルはneutralと出力していた
  - 提案モデルでは注目を集めることができ正解できた
- 2番目の例ではcuteがpositiveの事例に高頻度で出現していたせいで，文脈上関係ない事例でもpositiveと判定していた
  - 提案モデルでは文脈に無関係であることをアテンションが捉え正解できた

まとめ

- アテンションの注目量に基づいて自動で教師アテンションを作成する手法を提案
  - 最大の注目を集めた単語をactive/misleadingな単語として収集する
  - 以前のactive/misleadingな単語をマスクすることで低頻度語にも注目させる
- 収集した単語から作成した教師アテンションによって再学習することで性能向上を確認

感想

- フレーズの一部分やnotが絡む箇所でのアテンションが見てみたい
- 低頻度語のベクトル表現の質に影響を受けやすい？
  - たとえ低頻度語にアテンションを当てる下地が出来たとしても，それに対応するベクトルがあまり学習されておらず質が悪ければ意味がないのでは？

補足資料｜データセットのサイズ

Domain	Dataset	#Pos	#Neg	#Neu
LAPTOP	Train	980	858	454
	Test	340	128	171
REST	Train	2159	800	632
	Test	730	195	196
TWITTER	Train	1567	1563	3127
	Test	174	174	346