Assessing the Ability of Self-Attention Networks to Learn Word Order

Baosong Yang[†] **Longyue Wang**[‡] **Derek F. Wong**[†] **Lidia S. Chao**[†] **Zhaopeng Tu**[‡*]
[†]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.baosong@gmail.com, {derekfw,lidiasc}@umac.mo
[‡]Tencent AI Lab
{vinnylywang,zptu}@tencent.com

ACL読み会(2019/09/11)

紹介者: 中原 拓哉(名古屋大学 M1)

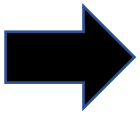
概要 ～ SANの語順学習能力の評価

- 研究動機
  - 再帰構造の欠如から，RNNに比べ，Self-Attention Networkでの語順学習能力は低いと推測されている
  - しかし，実験的にそれを確かめた例がなく，また，位置情報を欠いている場合の機械翻訳の強力なパフォーマンスの説明もなされていない
- 評価方法
  - Word Reordering Detection(WRD)タスクというものを独自に設定し，SANとRNN,DiSANで語順学習能力を比較する
- 結果
  - WRDの学習を行ったSANは，position encodingを行なったとしても，RNNほど語順情報を学習するのが困難
  - しかし，encoderを機械翻訳用のデータで学習させたSANでは，RNNよりも優れた語順情報を学習する

焦点を当てる3つの疑問と実験設定

疑問

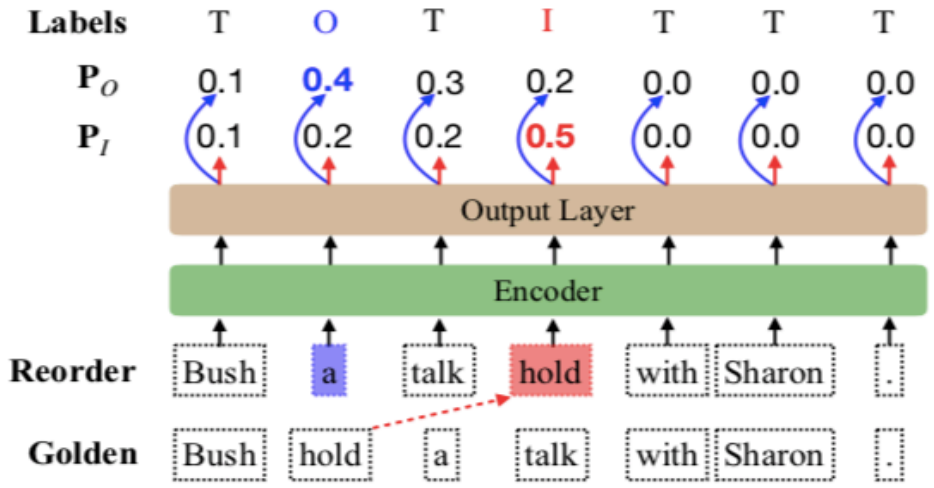
- **Q1.**
  - 再帰構造は語順学習に重要なものであるか？また，その結果は翻訳などの他のタスクにも当てはまるのか？
- **Q2.**
  - 機械翻訳のような下流のタスクにおける語順学習で，もっとも重要なのはモデル構造なのか？
- **Q3.**
  - SANが語順情報を得るために，position encodingを利用するのは十分なことなのか？



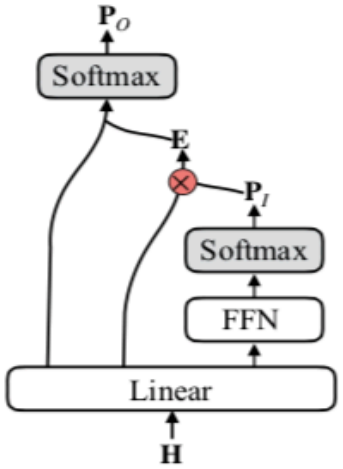
実験設定

- **Q1.**
  - WRDタスクにおけるSANと二つの再帰モデル(RNN，DiSAN)を比較して語順学習能力の定量化を行う
- **Q2.**
  - モデル構造と学習目標の効果を比較するために，各encoderをWRDとNMTそれぞれのデータで学習して比較
- **Q3.**
  - SAN, DiSANのposition encodingを取り除くことによって，position encodingの効果を評価

Word Reordering Detection(WRD)



(a) Position Detector



(b) Output Layer

$$\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_N\} \qquad \mathbf{U}_I \in \mathbb{R}^d$$

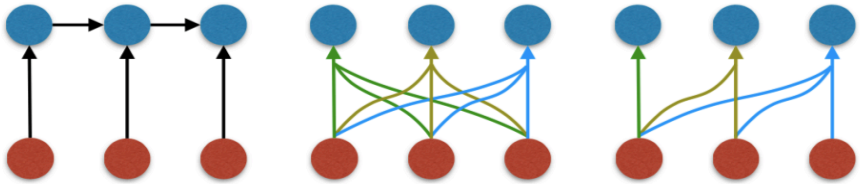
$$\mathbf{P}_I = \texttt{SoftMax}(\mathbf{U}_I^\top \texttt{tanh}(\mathbf{W}_I \mathbf{H})) \quad \in \mathbb{R}^N \qquad \mathbf{W}_I \in \mathbb{R}^{d \times d}$$

$$\mathbf{E} = \mathbf{P}_I(\mathbf{W}_Q \mathbf{H}) \quad \in \mathbb{R}^d \qquad \{\mathbf{W}_Q, \mathbf{W}_K\} \in \mathbb{R}^{d \times d}$$

$$\mathbf{P}_O = \textsc{Att}(\mathbf{E}, \mathbf{W}_K \mathbf{H}) \quad \in \mathbb{R}^N$$

$$L = \mathbf{Q}_I^\top \log \mathbf{P}_I + \mathbf{Q}_O^\top \log \mathbf{P}_O \qquad \{\mathbf{Q}_I, \mathbf{Q}_O\} \in \mathbb{R}^N$$

実験設定～Encoder Setting



- **RNN**

$$\mathbf{h}_n = f(\mathbf{h}_{n-1}, \mathbf{x}_n),$$

・連続的に隠れ状態を生成する.
・本論文において, fはGRUを表す.
・RNNは，前の状態に依存するので並列化が難しい.

- **SAN**

$$\mathbf{h}_n = \mathrm{ATT}(\mathbf{q}_n, \mathbf{K})\mathbf{V},$$

・並列構造で隠れ状態を生成する
・与えられたq(query)と各K(key)の類似度を測り, V(value)との加重和として取り出す
・Position encodingも利用している

- **DiSAN**

$$\mathbf{h}_n = \mathrm{ATT}(\mathbf{q}_n, \mathbf{K}_{\leq n})\mathbf{V}_{\leq n},$$

・語順をencodeする機能をSANに追加
・SANとは異なり, 自分より前の入力のみを用いる
・Position encodingも利用している

実験設定～学習目標とデータセット

- WRD Encoder (語順検出)
  - 最初からWRD用のdataでencoderを学習
  - output layerも一緒に学習
  - データセット
    - WMT14(En → De)のEnの1単語の位置を変更したもの(最大長80)
    - 最終的に，train : 7M , valid : 10K, test : 10K
- NMT Encoder (機械翻訳)
  - 最初に，NMTのdataでencoderを学習
  - その後，encoderのparameterは固定したまま，output layerをWRD用に学習
  - データセット
    - WMT14(En → De): 4.5M pairs , newstest2013 , newstest2014
    - WAT17(En → Ja): 2.0M pairs , newsdev2017, newstest2017

また，En,DeデータはMoses, JaデータはKeTeaによって単語分割
語彙数を減らすために，BPEも用いている

Results on WRD Encoders

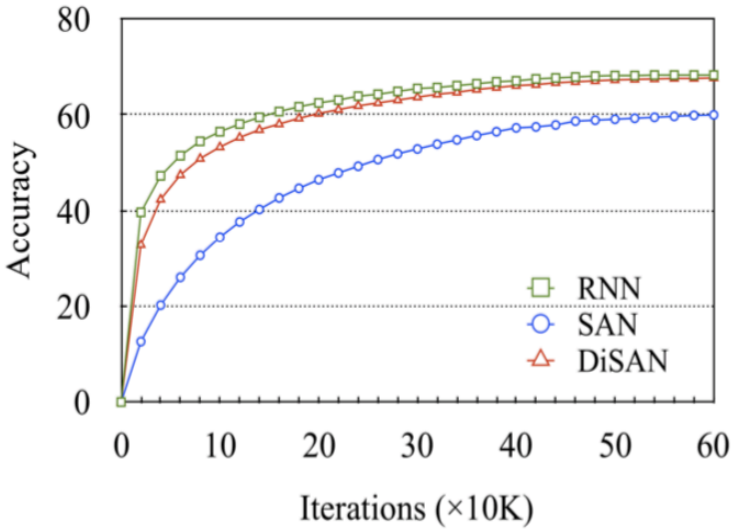
- WRDのAccuracyの結果
  - RNNとDiSANの両方がSANを超えるAccuracy

実際に，再帰構造を持つRNN,DiSANの方が語順情報に優れる

  - また，DiSANがSANよりもAccuracyが高いことから，WRDタスクの信頼性が高いことが示された

Models	Insert	Original	Both
RNN	78.4	**73.4**	**68.2**
SAN	73.2	66.0	60.1
DiSAN	**79.6**	70.1	68.0

- Learning abilityの結果
  - 他と比べるとSANの収束が遅く，並列構造が語順情報を学習しづらいという直感に一致する
  - DiSANは最初わずかにRNNよりも収束が遅いが，徐々に追いついていく



Results on Pre-Trained NMT Encoders

Model	Translation		Detection		
	En⇒De	En⇒Ja	En⇒De Enc.	En⇒Ja Enc.	WRD Enc.
RNN	26.8	42.9	33.9	29.0	**68.2**
SAN	27.3	43.6	**41.6**	**32.8**	60.1
- Pos_Emb	11.5	–	0.3	–	0.3
DiSAN	**27.6**	**43.7**	39.7	31.2	68.0
- Pos_Emb	27.0	43.1	40.1	31.0	62.8

- NMT(Translation)のAccuracyの結果
  - SANの方がRNNよりもAccuracyが高く，既存研究の結果と一致
  - さらに，DiSANの方がSANよりAccuracyが高いことから，モデルの方向情報が翻訳において有効
  - 以上のことから，モデルの信頼性が高いことを示す

Results on Pre-Trained NMT Encoders

Model	Translation		Detection		
	En⇒De	En⇒Ja	En⇒De Enc.	En⇒Ja Enc.	WRD Enc.
RNN	26.8	42.9	33.9	29.0	**68.2**
SAN	27.3	43.6	**41.6**	**32.8**	60.1
- Pos_Emb	11.5	–	0.3	–	0.3
DiSAN	**27.6**	**43.7**	39.7	31.2	68.0
- Pos_Emb	27.0	43.1	40.1	31.0	62.8

- WRD(Detection)のAccuracyの結果
  - SANベースのNMT encoder の方がAccuracyが高い
  - 特に，SANよりDiSANの方が翻訳のAccuracyが高いにも関わらず，WRDに関してはSANの方が高い

SANベースのencoderでも，機械翻訳の学習中では，語順に関する特徴を保持する

  - また，WRDタスクにおいて，全てのNMT encoderはWRD encoderよりもAccuracyが低い

語順をモデリングする上で，モデル構造よりも学習目標の方が重要？

Position Encoding VS. Recurrence Modeling

Model	Translation		Detection		
	En⇒De	En⇒Ja	En⇒De Enc.	En⇒Ja Enc.	WRD Enc.
RNN	26.8	42.9	33.9	29.0	**68.2**
SAN	27.3	43.6	**41.6**	**32.8**	60.1
- Pos_Emb	11.5	–	0.3	–	0.3
DiSAN	**27.6**	**43.7**	39.7	31.2	68.0
- Pos_Emb	27.0	43.1	40.1	31.0	62.8

Position encodingの有効性を確認するために，SANとDiSANからPosition encodingを取り除いたモデルで実験

- 結果
  - SANからPosition encodingを取り除くと，NMTでもWRDでも大幅にAccuracyが下がる
  - つまり，SANではPosition encodingの重要度が高い
  - DiSANのPos_EmbよりもSANの方がAccuracyが高い

再帰構造よりも，Position encodingの方が語順情報を学習するのに適している

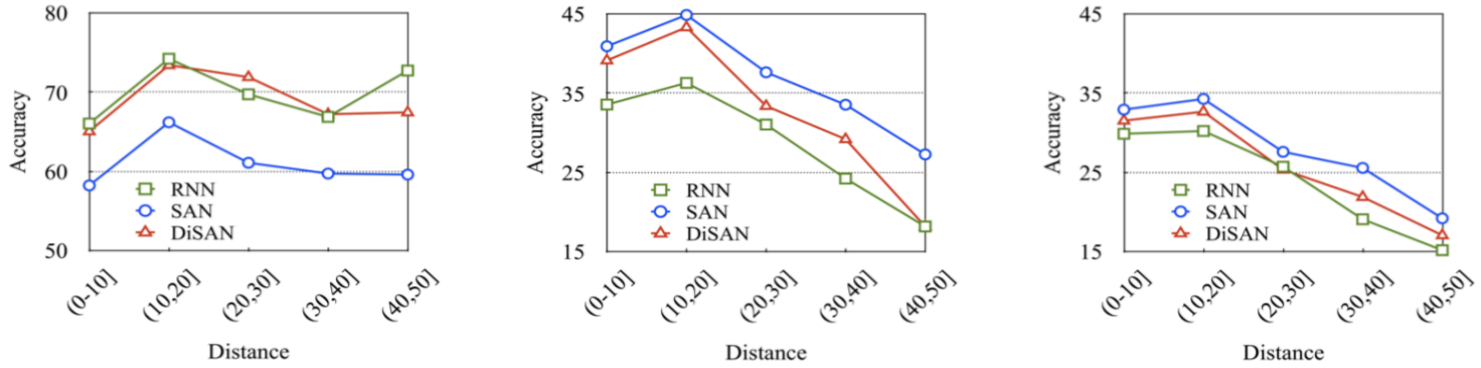
追加実験

1. 学習目標は本当に語順情報に影響するのか？

2. SANはposition encodingからどうやって語順情報を引き出しているのか？

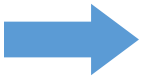
3. より多くの語順情報を保持することは，翻訳タスクに役立つことなのか？

Accuracy According to Distance



(a) WRD Encoder (b) En⇒De NMT Encoder (c) En⇒Ja NMT Encoder

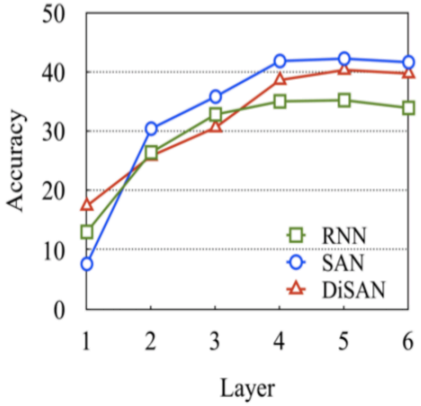
- 追加実験1
  - WRDタスクにおいて，pop位置とinsert位置の距離によるAccuracyの変化を調査
- 結果
  - WRD Encoderでは距離が遠くなるにつれて，わずかにAccuracyが低下
  - NMT Encoderでは距離が遠くなるにつれて，急激にAccuracyが低下
    - NMTが長距離システムが長距離依存に弱いという直感に一致
    - NMTでは長距離語順のような，ソース文を理解するのに比較的関係ない情報を破棄する傾向がある
    - しかし，WRDのようなNLPの下流タスクではそのせいで精度が落ちてしまう



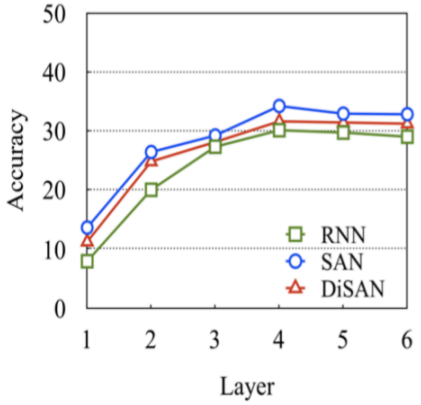
語順情報の学習に関しては，本論文のモデル構造よりも学習目標の方が，より影響する

Accuracy According to Layer

- 追加実験2
  - NMTタスクでどのように語順情報を学習しているのかを調べるために,
    NMT encoderの各レイヤーでのAccuracyを調査



(a) En⇒De NMT encoder



(b) En⇒Ja NMT encoder

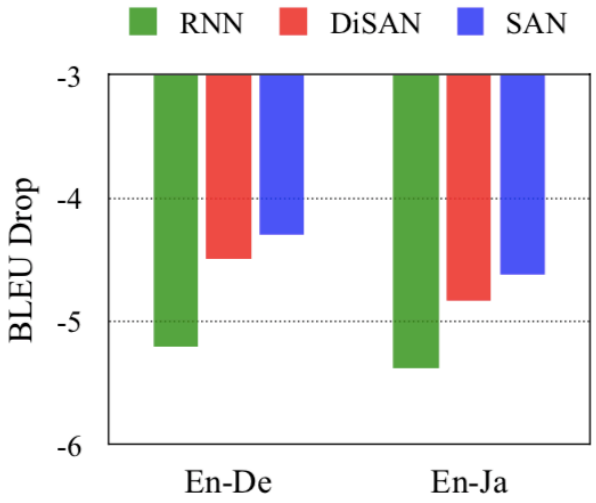
- 結果
  - ほぼ全てのレイヤーでSANのAccuracyが最も高い
  - 三つのモデル構造全てで，レイヤーが深くなるにつれAccuracyが上昇する

NMTタスクでは，言語特徴の学習とともに積み重ねるように語順情報をできるだけ保持しようとする

NMTタスクにおける語順に関して，position encodingが再帰構造と同じような学習効果を発揮する

Effect of wrong Word Order Noises

- 追加実験3
  - NMTモデルの学習において開発セット中に，ノイズとして1単語の語順誤りがあった場合の影響をBLEU dropで検証



- 結果
  - SANとDiSANはRNNよりもBLUE dropが小さい

語順誤りのノイズを除去することに対するself-attentionの有効性が示された

まとめ

- WRD(Word Reordering Detection)タスクを独自に設定し，RNN,SAN,DiSANにおける語順情報の学習能力を調査
- 調査結果
  - WRD encoderの場合はこれまでの直感通り，再帰構造を持つRNN,DiSANの方が，SANよりも語順情報を捉えることができる
  - しかし，NMT encoderにおいては，SANの方が語順情報を捉えられているっぽい
  - 語順などの具体的な特徴を学習する場合，学習目標がきわめて重要となることがある
  - NMTタスクにおいて，RNNは語順誤りによるノイズの影響が大きい
- NMT encoderに限らず，他のNLPタスクで訓練したencoderでも語順情報をどのように学習しているのかを実験することに興味を示している