英日翻訳タスクにおけるスワップモデルを通した seq2seq と Transformer の比較

根石将人（東京大学）　吉永直樹（東京大学生産技術研究所）
neishi@tkl.iis.u-tokyo.ac.jp　ynaga@iis.u-tokyo.ac.jp

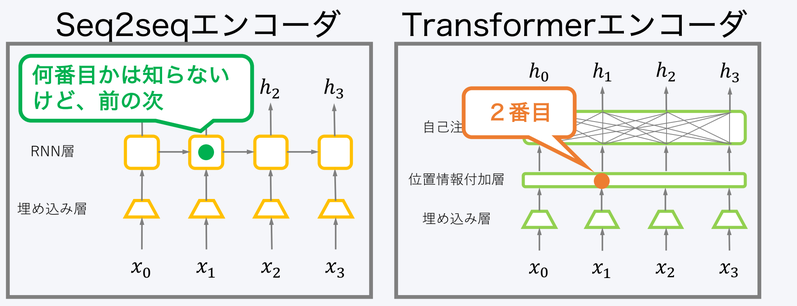
背景： 様々な構造のニューラル機械翻訳 (NMT) モデルが提案され、翻訳精度が向上している一方、向上幅は減少傾向にある

目的： 基本的なNMTモデルであるseq2seq[1]とTransformer[2]の比較を通して、NMTモデルの改善余地を探る

方法： NMT (seq2seq) の課題（学習データ量、ビームサーチ、ドメイン外翻訳、文長、低頻度語、単語アラインメント）を統計的機械翻訳との比較を通して指摘したKoehnら[3]の研究を参考に、モデル構造に注目しつつ、スワップモデルを導入しエンコーダ・デコーダ単位で、4つの課題について比較分析を行う

Seq2seq と Transformer の主な構造的な違い (エンコーダ・デコーダ共通)

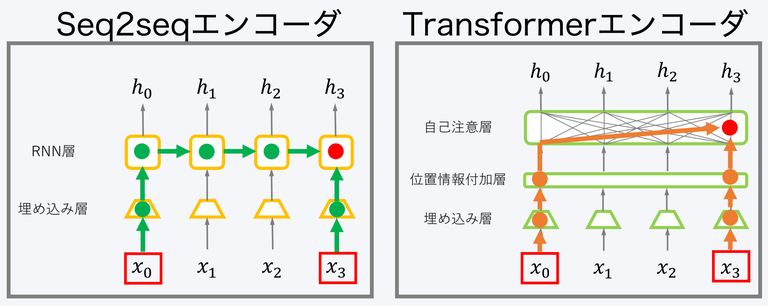
入力系列 (文) の位置の扱い



☺RNNの逐次処理の中で**相対的**に扱われる

☹**絶対**位置に基づくベクトルを単語ベクトルに加算するが、この位置ベクトルの学習が必要になる

単語間のネットワーク内距離



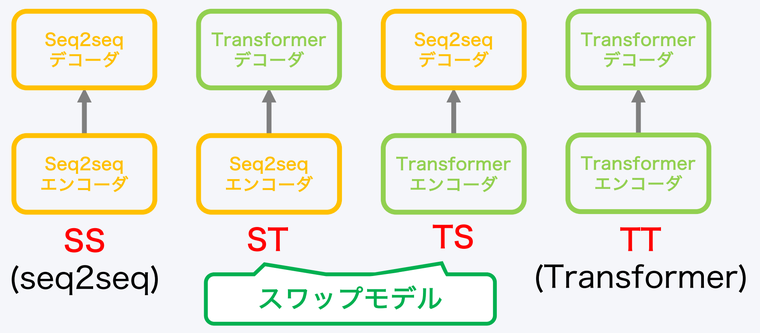
☹先に入力された単語の情報は徐々に失われ、離れた単語間の関係は捉えにくい

☺自己注意層により、全ての単語間の距離は等しく短く、離れた単語間の関係も扱える

☺密なネットワークは学習(誤差伝搬) にも貢献

スワップモデルの導入

スワップモデル
- Chenら[4]のhybrid modelの一種
- 2つのモデルのエンコーダとデコーダを入れ替える



エンコーダ・デコーダ単位での細かい分析を可能に

実験

実験設定

■ データセット (英日翻訳)

語彙数：16,000 (英日共通)
* SentencePiece (unigram) [Kudo+, 2018] による単語分割

対訳コーパス	学習データ量（前処理後）	ジャンル
ASPEC	1.3M	科学技術論文
KFTT	0.4M	京都関連 wikipedia
JESC	3.1M	映画、TVの字幕

■ モデル
- Seq2seq　[エンコーダ：3層双方向LSTM、デコーダ：1層LSTM]
- Transformer [エンコーダ、デコーダ：6層]

全て、単語埋め込み、隠れ層の次元数は512

参考：モデルのパラメータ数

	エンコーダ	デコーダ
seq2seq	28M	21M
Transformer	27M	41M

■ 学習設定
最適化手法：Adam (初期学習率0.0001)、バッチサイズ：128、学習ステップ：400k
※KFTTは64

■ 評価尺度：BLEU [Papineni+, 2002]

0. 基本的な翻訳精度の確認

	ASPEC	KFTT	JESC
SS	*37.02*	*28.39*	**16.60**
ST	**39.19**	30.20	**17.24**
TS	37.74	29.08	*16.14*
TT	38.65	**31.04**	16.36

*太字は、ブートストラップ検定において、最高スコアモデルに対するp値が0.05以上 (同等に良い)

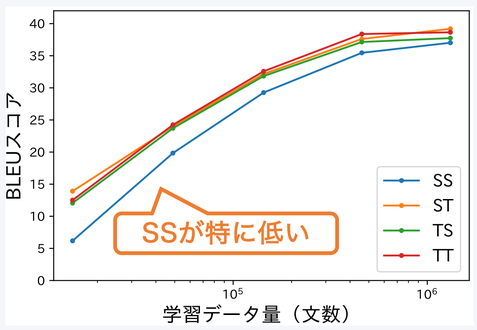
ST >= TT > TS > SS

エンコーダ：S ≒ T
デコーダ：　S < T

1. 少ない学習データ量で効率的に学習できるか？

データセット：ASPEC

学習データの先頭n文を用いて学習
n = 15k, 50k, 150k, 500k, 1.5M
*ここから更に文長50で制限



エンコーダ、デコーダが共にseq2seqのモデルのみスコアが低い

逆に**どちらかのみ**でもTransformerであれば翻訳精度向上

2.ビームサーチ時のビーム幅は翻訳精度にどう影響するか？

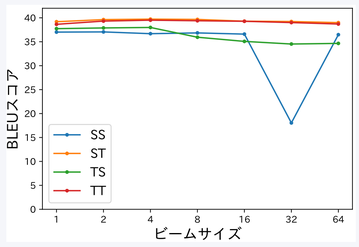
データセット：ASPEC

ビーム幅を1~64まで変化させて開発データで評価

STとTTがほぼ同じ安定した挙動

Tデコーダがより安定かつ効果的だが、有意差は認められない

開発データでの翻訳精度



NEW

参考：テストデータでの翻訳精度

	BLEU	Δ	ビーム幅
SS	37.05	*+0.03*	8
ST	39.68	**+0.49**	4
TS	38.00	+0.26	4
TT	39.02	+0.37	4

3.学習ドメイン外テキストの翻訳におけるモデルの頑健性に違いはあるか？

それぞれのデータセットで学習したモデルを、それぞれのテストデータで評価

学習データ	ASPEC			KFTT			JESC		
テストデータ	ASPEC	KFTT	JESC	ASPEC	KFTT	JESC	ASPEC	KFTT	JESC
SS	*37.02*	7.08	1.81	*8.11*	*28.39*	2.32	4.25	**3.75**	16.60
ST	**39.19**	**8.09**	2.09	9.75	30.20	3.39	**4.68**	3.68	**17.24**
TS	37.74	*5.02*	*0.90*	8.48	29.08	*2.29*	*3.11*	3.20	*16.14*
TT	38.65	7.99	**2.43**	**10.26**	**31.04**	**3.70**	3.43	*2.48*	16.36

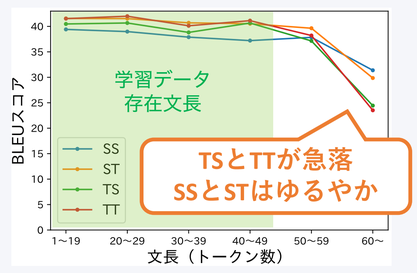
モデル構造の違いによる特別な改善は見られない

4. 学習データにない文長は扱えるか？

データセット：ASPEC

実験1. テストデータを原言語文の文長毎に分割して評価　* 学習データは文長50で制限

学習データ外の長文でTSとTTのスコアが急落
Sエンコーダ > Tエンコーダ



実験2. 学習データを原言語文の文長毎に分割して各モデルを学習
テストデータを原言語文の文長毎に分割して評価

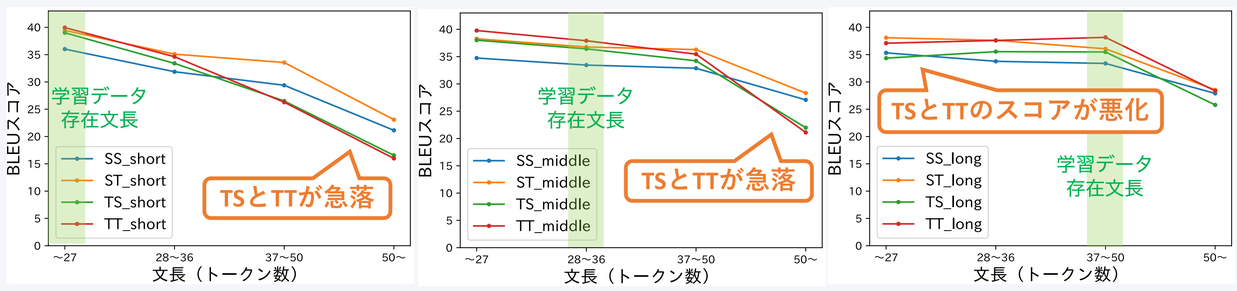
NEW

学習データの分割の内訳

	最小文長	最大文長	文対数	単語数
short	2	27	623,520	12,327,165
middle	27	36	394,949	12,327,168
long	36	50	296,026	12,327,117

学習データがある：
Sエンコーダ < Tエンコーダ

学習データがない：
Sエンコーダ > Tエンコーダ



Long 学習データ で学習したモデルの翻訳例

原文	the merit and demerit of Intranet are explained .
参照訳	イントラネットのメリットとデメリットを解説した。
SS	イントラネットの長所と短所を説明した。
ST	イントラネットのメリットとデメリットを説明した。
TS	イントラネットの長所と短所を説明した。
TT	イントラネットの長所と短所を説明した後,イントラネットの長所と短所を説明した。

絶対位置の場合、長文だけでの学習では、出力文が短い時点での終了が困難に（平均出力長も長い）

まとめ

- エンコーダはseq2seqとTransformerのどちらも同等の翻訳精度を実現する(実験0)が、Transformerの位置ベクトルの影響により、文長について学習データがある場合はTransformerが、無い場合はseq2seqが優れる(実験4)
- デコーダはTransformerが優れ(実験0)、ビームサーチでも安定した効果が得られる(実験2)
- エンコーダ・デコーダどちらかのみでも、Transformerの密なネットワークは翻訳精度を向上させる(実験1)
- モデル構造の工夫では、学習ドメイン外の翻訳は解決できていない(実験3)

今後の予定

- 文長について、Tエンコーダの位置ベクトル学習のために、複数文を繋ぐなど、長文を生成するデータ拡張を試みる
- 現状改善のない学習ドメイン外テキストの翻訳問題の解決

参考文献

[1] Luong+, Effective approaches to attention-based neural machine translation, EMNLP 2015
[2] Vaswani+, Attention is all you need, NIPS 2017
[3] Koehn+, Six Challenges for Neural Machine Translation, ACL 2017
[4] Chen+, The best of both worlds: Combining recent advances in neural machine translation, ACL 2018