目的言語側の文間文脈を考慮した文脈つきニューラル機械翻訳

首都大学東京　山岸駿秀　小町守　yamagishi-hayahide@ed.tmu.ac.jp

概要

近年のニューラル機械翻訳（NMT）では、扱える文脈を文間文脈へ拡張する動きがある。現在提案されている手法は主に対象文の1つ前の入力文を用いるものであるが、原言語側に比べて目的言語側の有用性は低いとされている。本研究では、その理由として目的言語側の情報をEncoderの枠組みで扱っていることが問題であると考え、目的言語側の文をDecoderの枠組みで扱った場合にどのような影響があるかについての比較検討を行った。

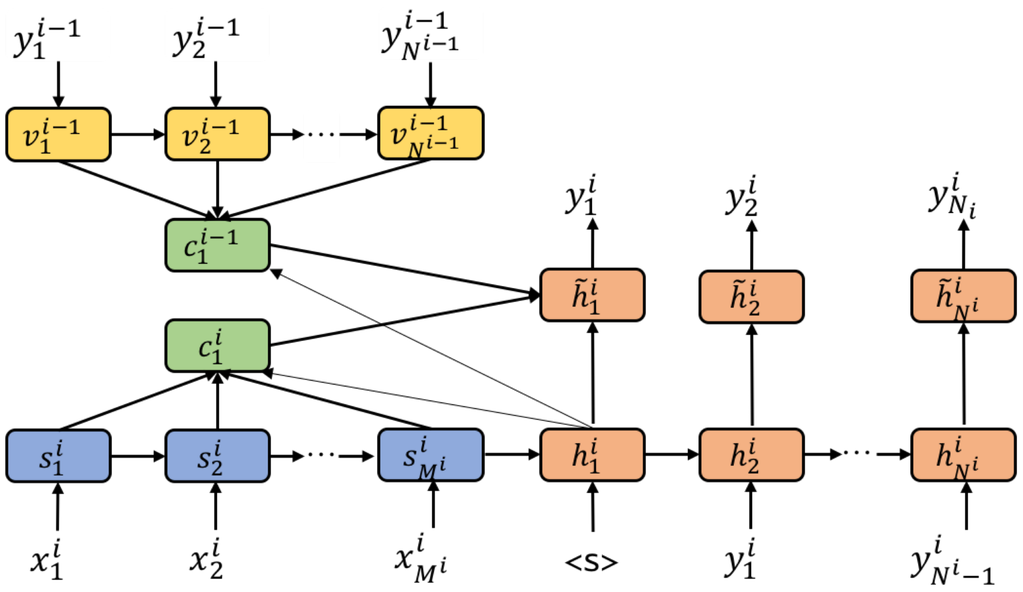
Multi-Encoder型の文脈つきニューラル機械翻訳

[Bawden+, NAACL18] のモデルをベースに文脈つきNMTを実装
・i文目の生成時にi-1文目の情報を得る文脈Encoderを用意し、その隠れ層ベクトルに対してもAttention c_n^{i-1}を計算
・ネットワークの式は以下のとおり（Zには X^{i-1} またはY^{i-1}が入る）

$$P(Y^i \mid X^i, Z) = \prod_{n=1}^{N^i} p(y_n^i \mid y_{<n}^i, X^i, Z), \quad p(y_n^i \mid y_{<n}^i, X^i, Z) = \mathrm{softmax}(W_o \tilde{h}_n^i), \quad \tilde{h}_n^i = \tilde{W}[h_n^i; c_n^i; c_n^{i-1}]$$

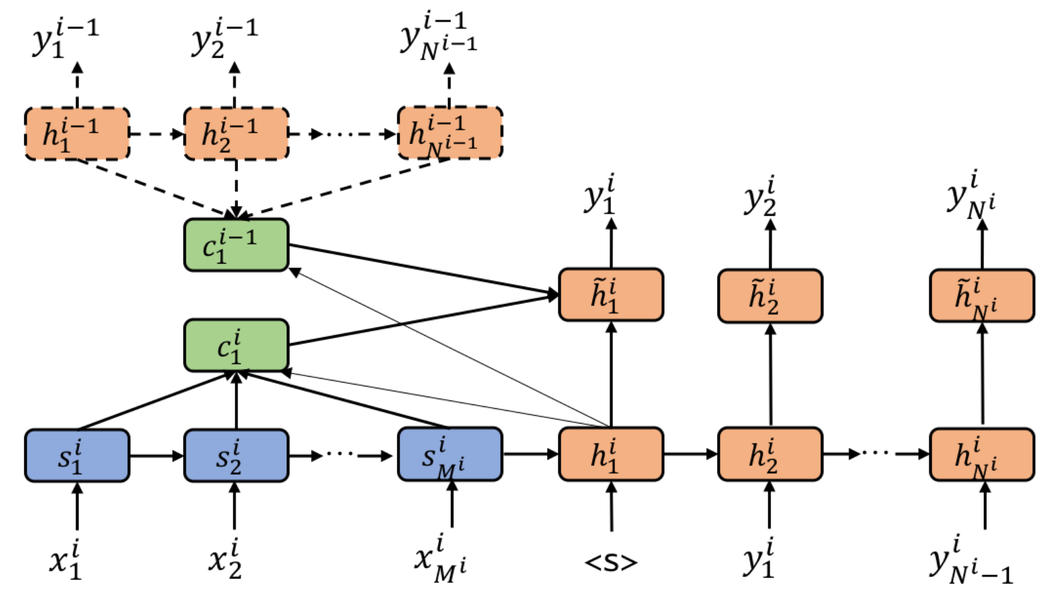
Separated型　[Bawden+, NAACL2018]など

1文前の情報を取り込む文脈Encoderを新たに用意
→ **入出力の情報をEncoderからの情報**として扱う



Shared型（提案手法）

1文前の翻訳時の隠れ層ベクトルを保存しておき、それを文脈Encoderの代わりに使う
→**入出力の情報をEncoder/Decoderの情報**として扱う



実験内容

以下の6モデルを比較
・Baseline: Dot型の Global Attention [Luong+, EMNLP15]
・Separated Source: 前の**入力文をEncode**する
・Separated Target: 前の**出力文をEncode**する
・Shared Source: 前の**生成時のEncoder**から情報を取得
・Shared Target: 前の**生成時のDecoder**から情報を取得
・Shared Mix: SourceとTargetのAttentionの和を使用

・Baselineと共通部分はPretrainしたものを初期値とする
・異なる乱数で**3回試し、BLEUの平均と標準偏差**を報告

ハイパーパラメータとデータ

Hidden / Embedding size	512
Batch size	128文書
Optimizer	AdaGrad（LR = 0.01）
学習	30 Epoch

コーパス	BPE	Train	Dev	Test
TED 独英/英独	32,000	203,998	888	1,305
TED 中英/英中	32,000	226,196	879	1,297
TED 日英/英日	32,000	194,170	871	1,285
Recipe 日英/英日	8,000	108,990	3,303	2,804

実験結果

・Shared Target型がほぼ全ての言語対で最高性能
→ 目的言語側の情報はDecoder側から得るべき

・パラメータ数は少ないが、Shared Source型も高性能
→ 重み共有自体に意味がある？ [Dabre+, AAAI19]
→ マルチタスク学習をしている？

・ドメインの違いは無関係らしい

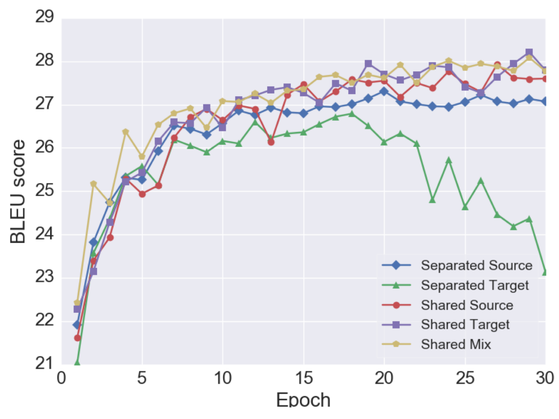
	Baseline	Separated Source	Separated Target	Shared Source	Shared Target	Shared Mix
TED 独英	26.55	26.29±.37	26.53±.12	*27.20±.11	***27.34±.11**	27.18±.21
TED 英独	21.26	21.04±.64	20.77±.10	21.63±.27	**21.83±.30**	21.50±.29
TED 中英	12.54	12.52±.33	12.63±.24	*13.36±.41	***13.52±.10**	*13.23±.09
TED 英中	8.97	8.94±.11	8.71±.06	9.45±.22	***9.58±.13**	9.42±.19
TED 日英	5.84	*6.64±.26	*6.37±.12	*6.95±.07	***6.96±.18**	*6.81±.16
TED 英日	8.40	8.58±.12	8.26±.00	8.51±.31	8.59±.08	**8.66±.14**
Recipe 日英	25.34	*26.51±.09	*26.69±.15	*26.90±.17	***26.92±.10**	*26.78±.11
Recipe 英日	20.81	*21.87±.12	*21.45±.14	***22.02±.20**	*21.97±.09	*21.81±.15

考察

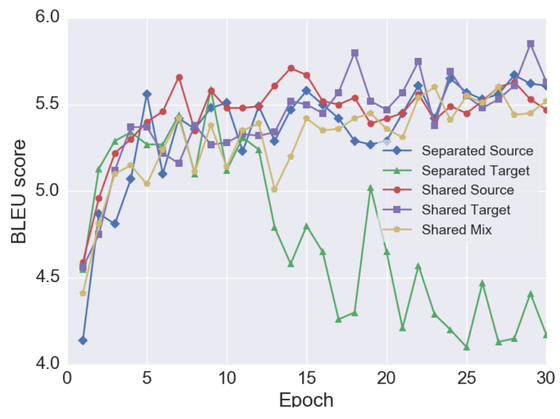
・Separated Target型の学習だけ不安定
→ Decoderと文脈Encoderの両方でExposure biasの影響？
→ Shared型は影響を減らせる

・日本語の有無で傾向が変化
→ 省略や語順等が関係している？

・一部の出力例では訳語の一貫性の改善を確認
　・ほとんどの例は全体的な質の向上だった



TED**独英**のDevのBLEU推移



TED**英中**のDevのBLEU推移