第233回自然言語処理研究会　2017/10/24

ニューラル機械翻訳における埋め込み層の教師なし事前学習

東京大学大学院情報理工学系研究科

根石将人　佐久間仁　遠田哲史　石渡祥之佑

東京大学生産技術研究所

吉永直樹　豊田正史

大規模ニューラルネットワークは パラメタの最適化が難しい

事前学習：データの量が不十分なタスクの学習

1. 別のドメインの豊富なデータで汎用的な学習
2. 目的のデータでタスク特化の学習（fine-tuning）

NMT（ニューラル機械翻訳）では

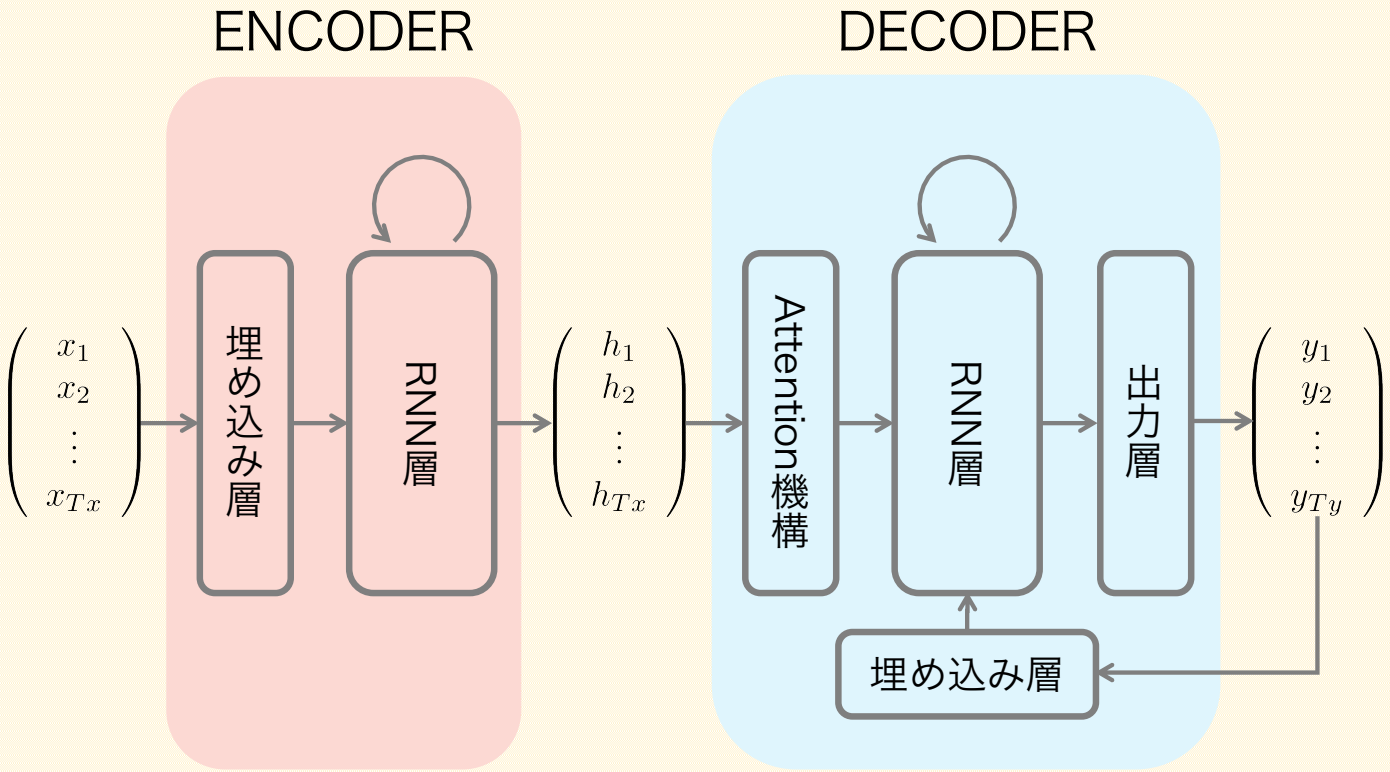
比較的少ない対訳コーパスに対して、

大規模単言語コーパスで言語モデルを学習し

- NMTモデルに統合
- パラメタの初期化（事前学習）

Attention機構付き Encoder-Decoderモデル

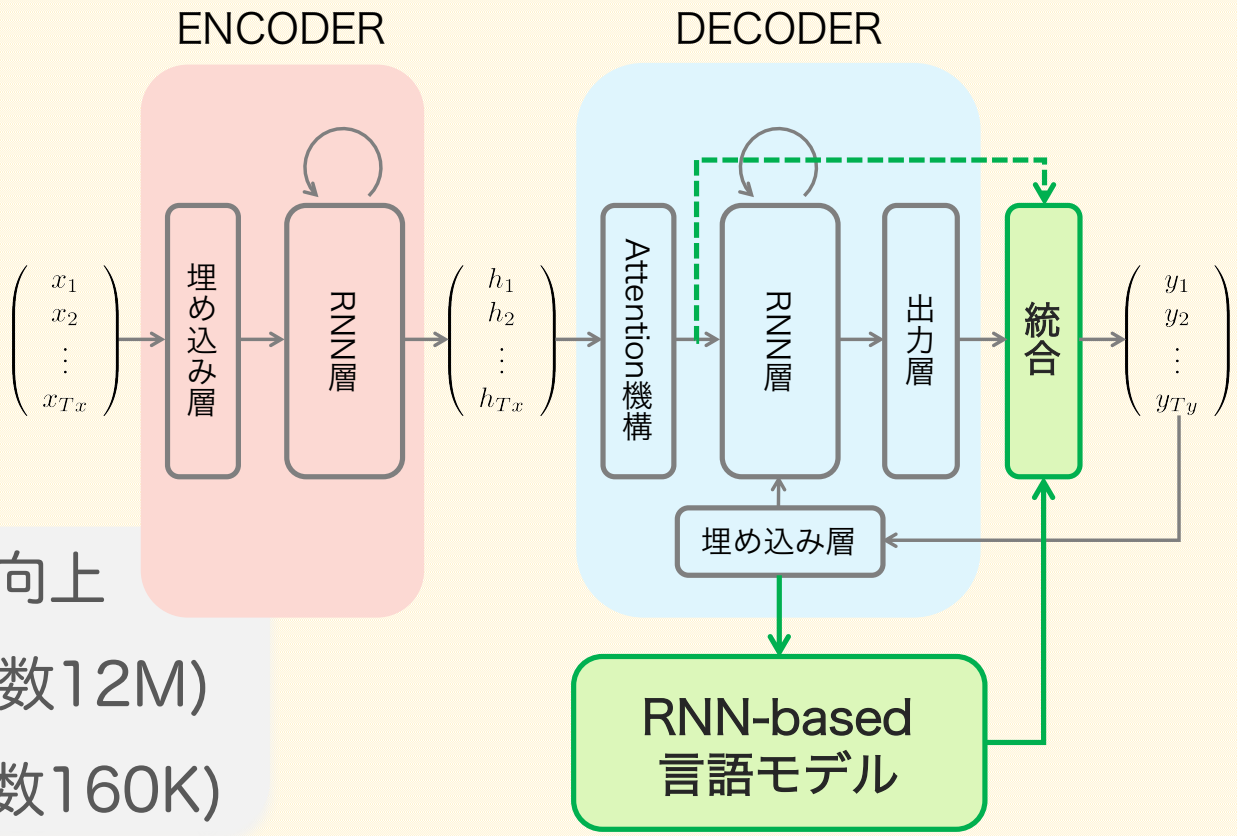
- NMTで主流のNNモデル
- EncoderとDecoderの２つのRNNから構成される



先行研究(1/3)：NMTと言語モデルの統合

Gulcehreら（2015）（事前学習ではない）

対訳コーパスが不十分な言語対の翻訳において、
目的言語(En)の大規模単言語コーパス(26GB)で学習した
言語モデルをNMTモデルに統合



BLEUスコアの向上

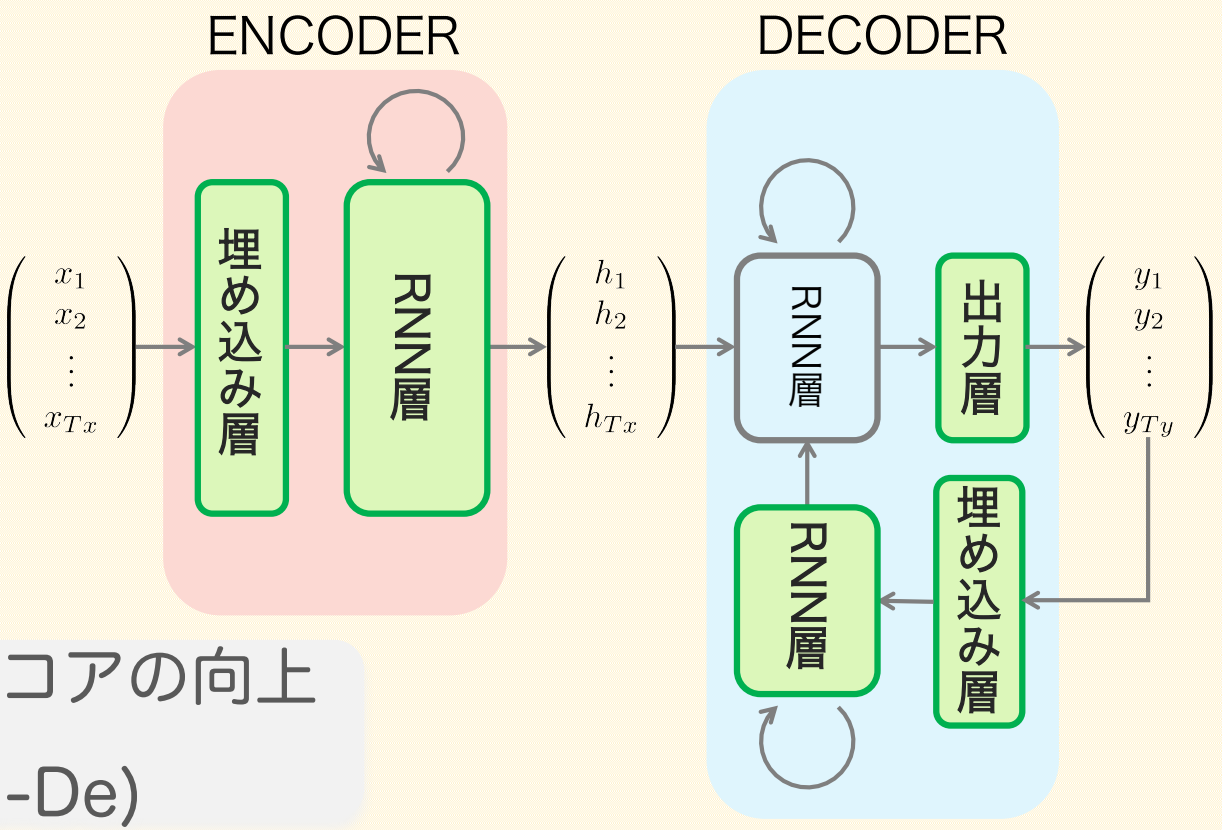
0.39(Cs-En:文数12M)

1.96(Tr-En:文数160K)

先行研究(2/3)：言語モデルによる初期化

Ramachandranら（2017）（NMTかつ事前学習）

- 両言語で、大規模単言語コーパス(4GB*2)で学習した言語モデルによる埋め込み層とRNN層、出力層の初期化
- fine-tuning時に言語モデルも同時学習し過学習を防止



BLEUスコアの向上

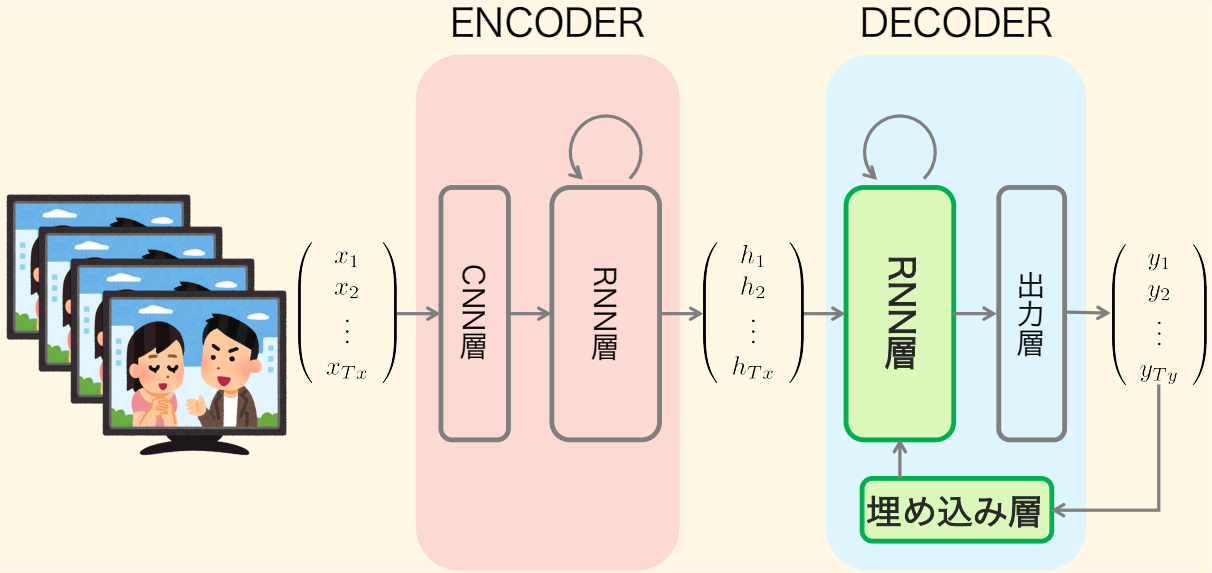
2.70(En-De)

先行研究(3/3)：GloVeによる初期化

Venugopalanら（2016）（NMTではない）

動画のキャプション生成タスクにおいて、

- （Gulcehreらを踏まえた異なる統合モデル）
- 大規模単言語コーパスで学習した言語モデルによる埋め込み層とRNN層の初期化
- 大規模単言語コーパス(24GB)とGloVe[3]による埋め込み層の初期化



[3]Pennington et al., 2014

着想

Venugopalanら（2016）

大規模単言語コーパスを使ってGloVeで事前学習した

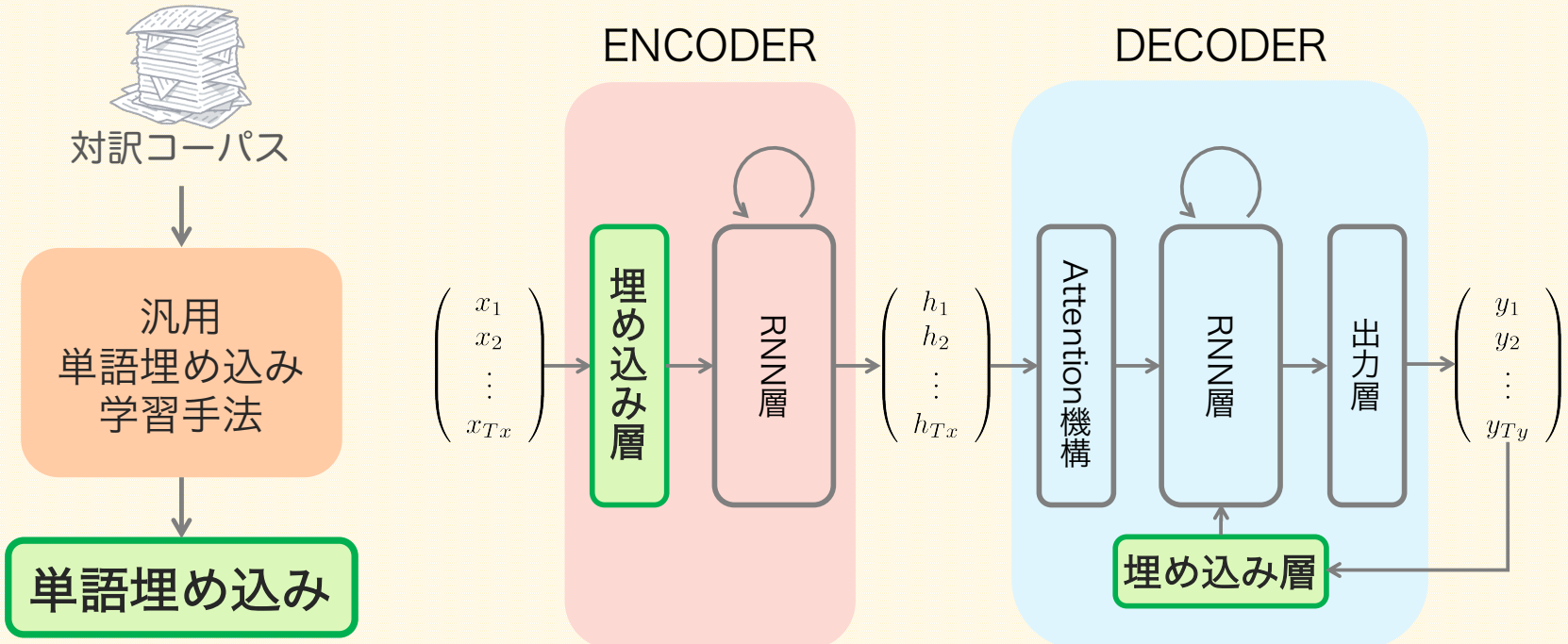
単語埋め込みによる埋め込み層の初期化が効果があった

- 埋め込み層のみの事前学習で十分効果があるのでは
- 単語埋め込みの学習はドメインを絞った方が良いのでは

  →対訳コーパスだけ

本研究の提案：
埋め込み層の教師なし事前学習

対訳コーパスのみを用いて、
汎用単語埋め込み学習手法（CBOWなど）で事前学習した単語埋め込みで、
NMTモデルの埋め込み層を初期化する。



実験リスト

主実験

- 埋め込み層のCBOWによる初期化（対訳コーパスのみ、+大規模コーパス）とランダム初期化の比較

追加実験

- 単語埋め込みの事前学習方法の比較
- 初期化する2つの埋め込み層の影響の比較
- 初期化後の埋め込み層のパラメタ固定

基本実験設定

- ASPEC日英対訳コーパスを用いた英日翻訳タスク
  (WAT2017のタスク)

- Attention機構付きのEncoder-Decoderモデル
  Encoder：2層の双方向LSTM
  Decoder：4層のLSTM

- SentencePiece（サブワードの拡張）
  両言語混在の語彙（共通トークンの存在による）

- ニューラル言語モデル
  CBOW（窓幅：5）

実験リスト

主実験

- 埋め込み層のCBOWによる初期化（対訳コーパスのみ、+大規模コーパス）とランダム初期化の比較

追加実験

- 単語埋め込みの事前学習方法の比較
- 初期化する2つの埋め込み層の影響の比較
- 初期化後の埋め込み層のパラメタ固定

CBOWによる埋め込み層初期化(1/4)

- 埋め込み層の初期化方法の比較

(1) ランダム初期化

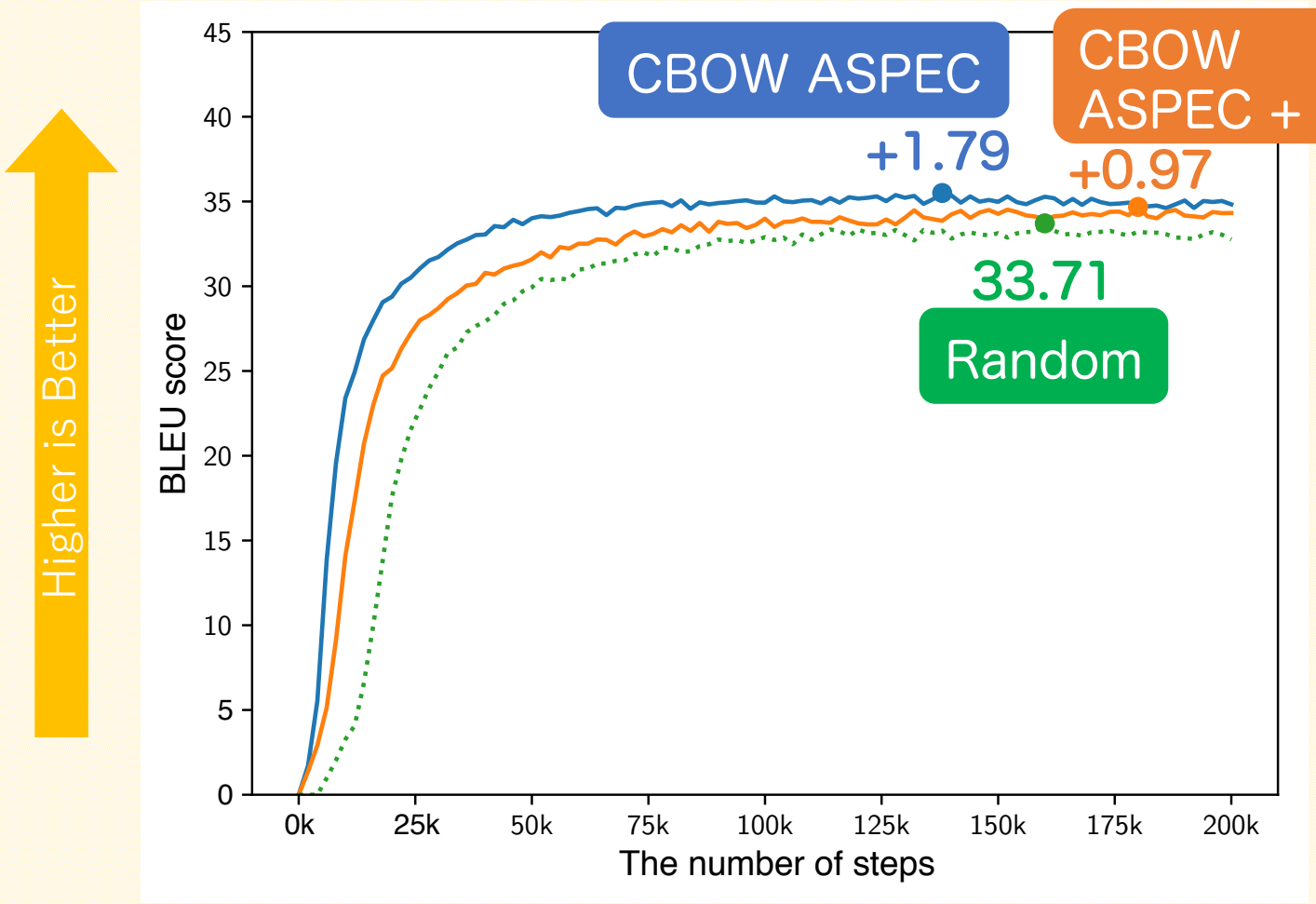
(2) CBOWによる初期化

ASPEC対訳コーパスのみ（400MB*2）

(3) CBOWによる初期化

ASPEC + 大規模コーパスWikipedia(12GB+4GB)

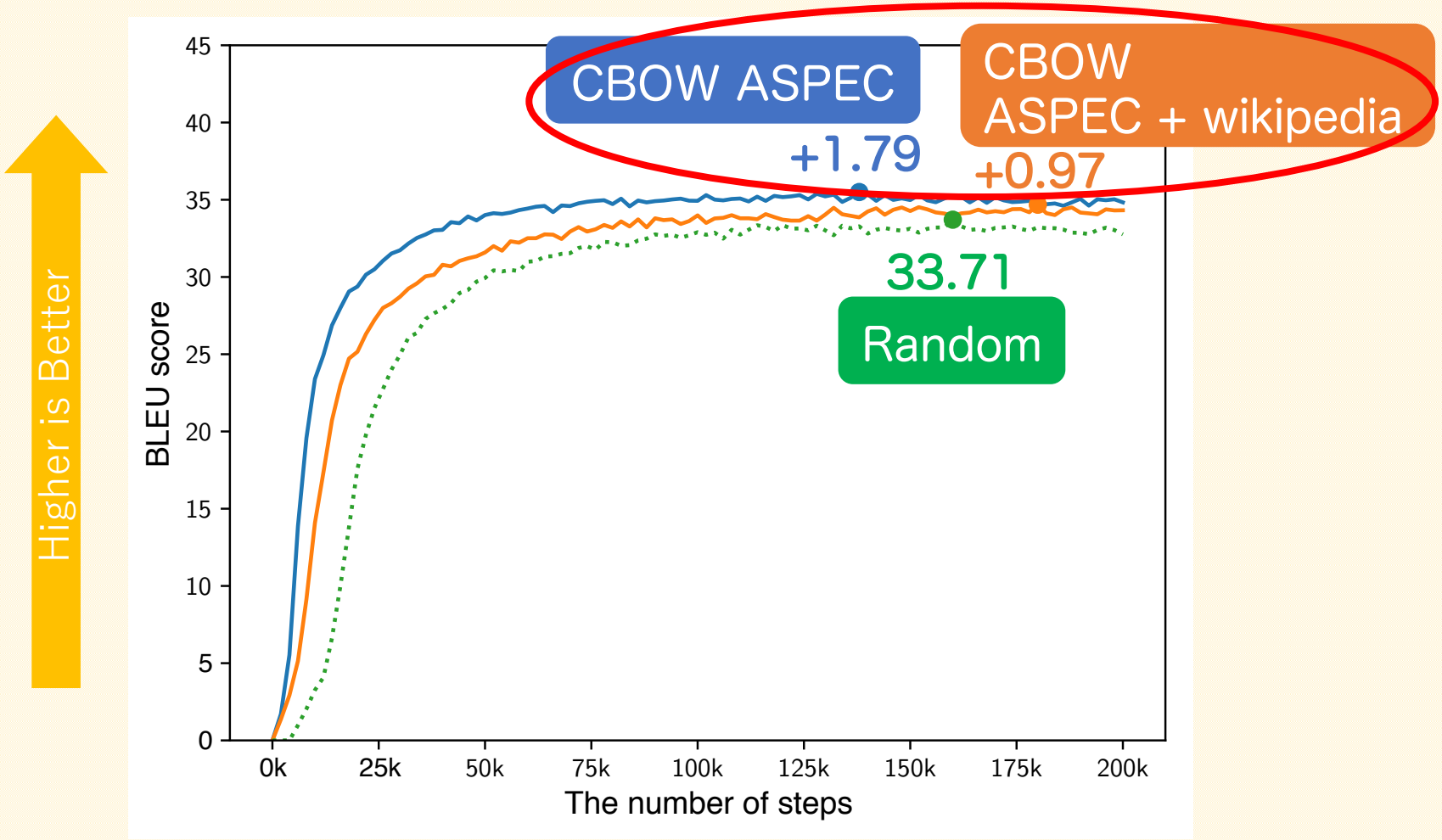
CBOWによる埋め込み層初期化(2/4)



それぞれの学習曲線

BLEUスコアが高いほど良い

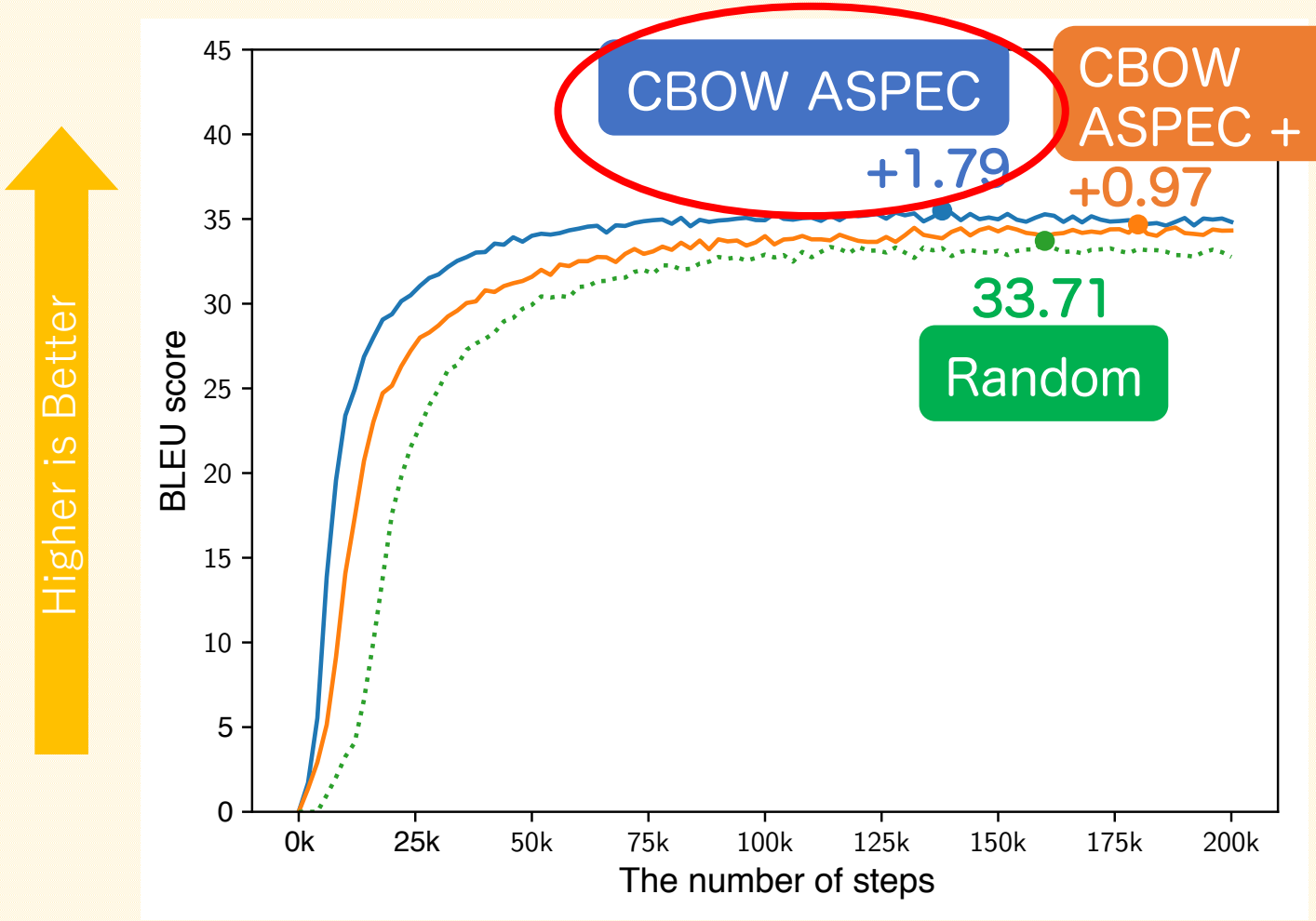
CBOWによる埋め込み層初期化(3/4)



CBOWによる初期化によって、収束解だけでなく

どのステップにおいてもBLEUスコアが向上した。

CBOWによる埋め込み層初期化(4/4)



大規模コーパスを追加せずに、

対訳コーパスのみの場合の方がより良い結果となった

実験リスト

主実験

- 埋め込み層のCBOWによる初期化（対訳コーパスのみ、+大規模コーパス）とランダム初期化の比較

追加実験

- 単語埋め込みの事前学習方法の比較
- 初期化する2つの埋め込み層の影響の比較
- 初期化後の埋め込み層のパラメタ固定

単語埋め込みの学習手法の影響(1/2)

- どの単語埋め込みの学習手法が適しているか

手法：CBOW[1]、Skip-gram[1]、SI-Skip-gram[2]、GloVe[3]

窓幅：2, 5, 10, (15 for GloVe)

- 追加実験

  埋め込み層のみについて、一度学習済みのNMTモデルのパラメタを初期値とし再学習

学習済みモデルの一度目の学習での初期化方法：

  ランダム初期化、CBOW(5)

[1]Mikolov et al., 2013 [2]Bojanowski et al., 2017 [3]Pennington et al., 2014

単語埋め込みの学習手法の影響(2/2)

- どの単語埋め込みの学習手法が適しているか

全ての手法において窓幅5が最高スコア

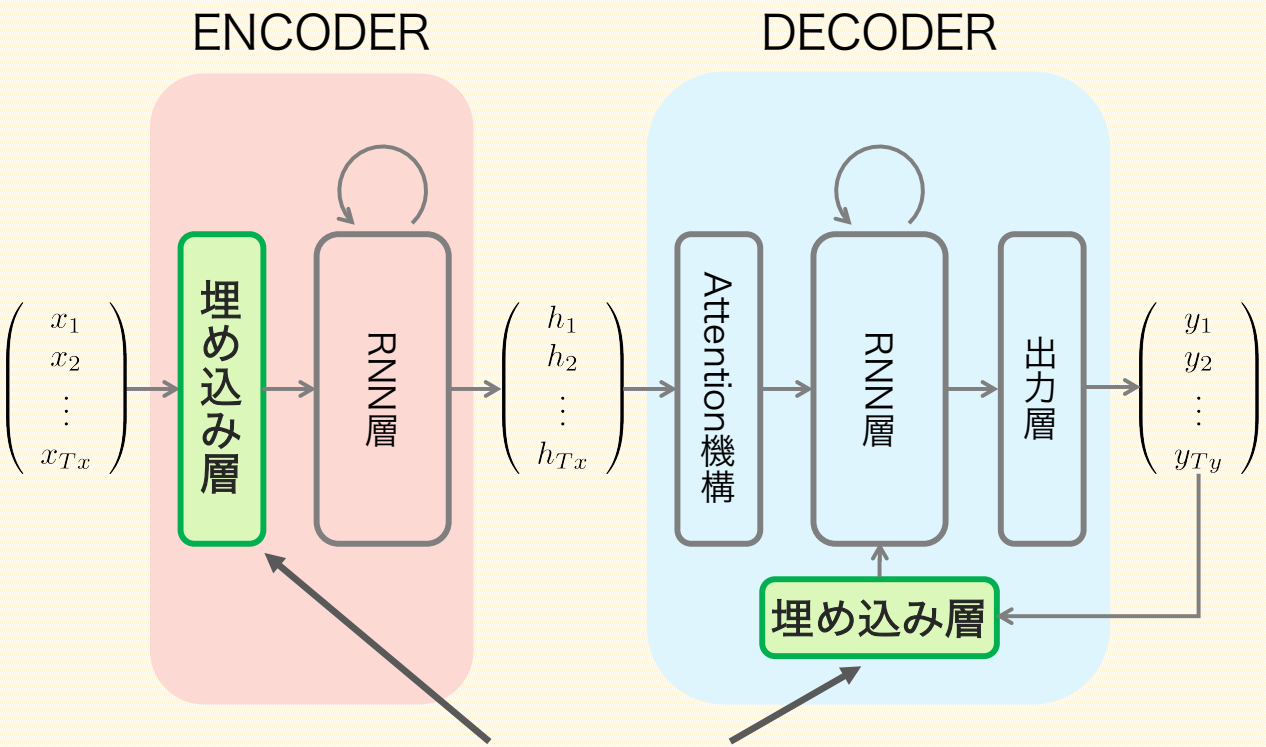
初期化手法	BLEUスコア	スコア差
ランダム	33.71	0
CBOW	**35.50**	**+1.79**
Skip-gram	34.44	+0.73
SI-Skip-gram	34.44	+0.73
GloVe	34.58	+0.77
学習済みモデル ランダム	33.81	+0.10
学習済みモデル CBOW	35.14	+1.43

窓幅5のCBOWが最良

学習済みモデルを用いた再学習は効果なし

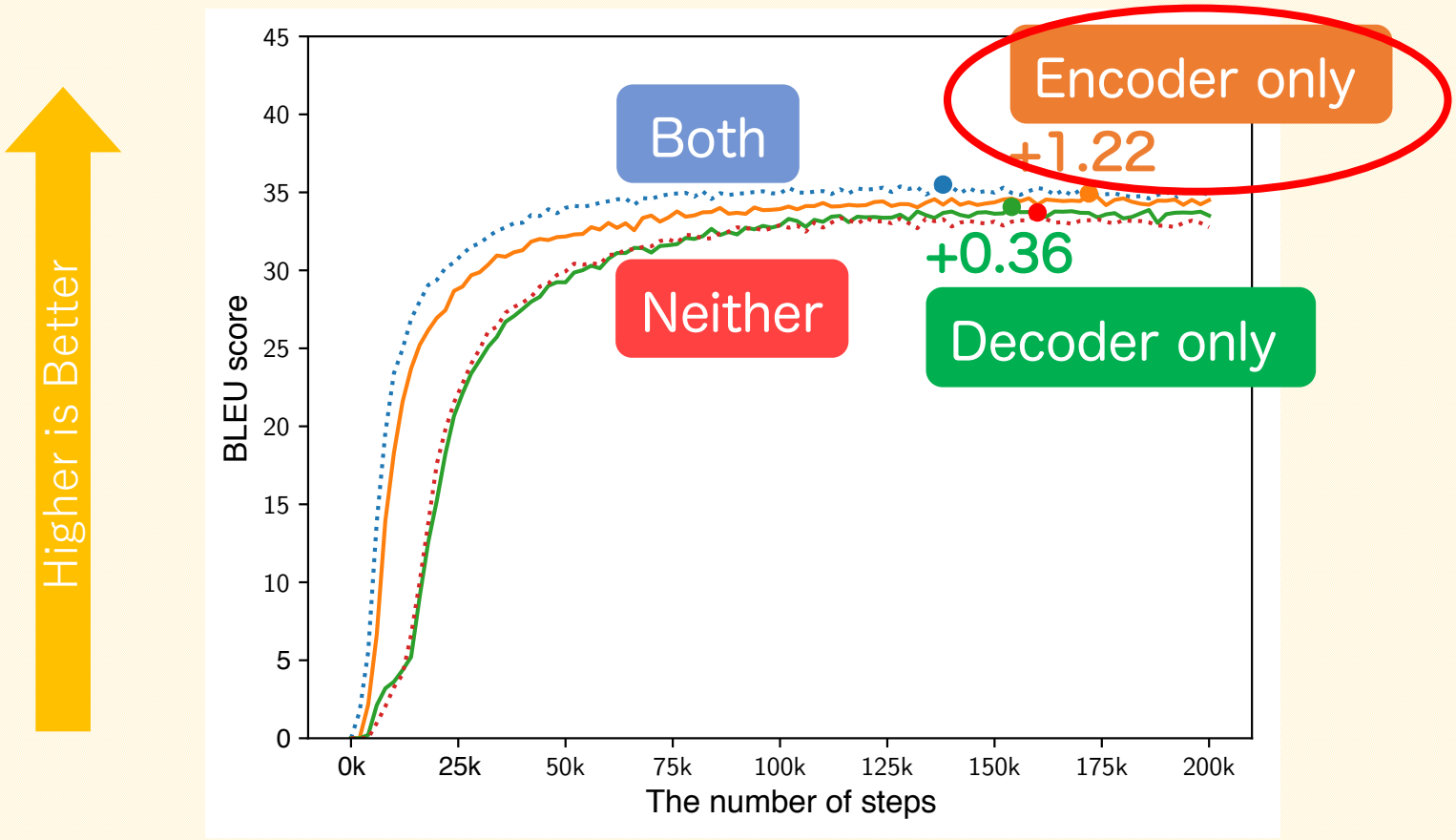
EncoderとDecoderの比較(1/2)

- EncoderとDecoderの2つの埋め込み層について、CBOWによる初期化をどちらかのみに限り、それぞれの影響を検証



EncoderとDecoderの比較(2/2)

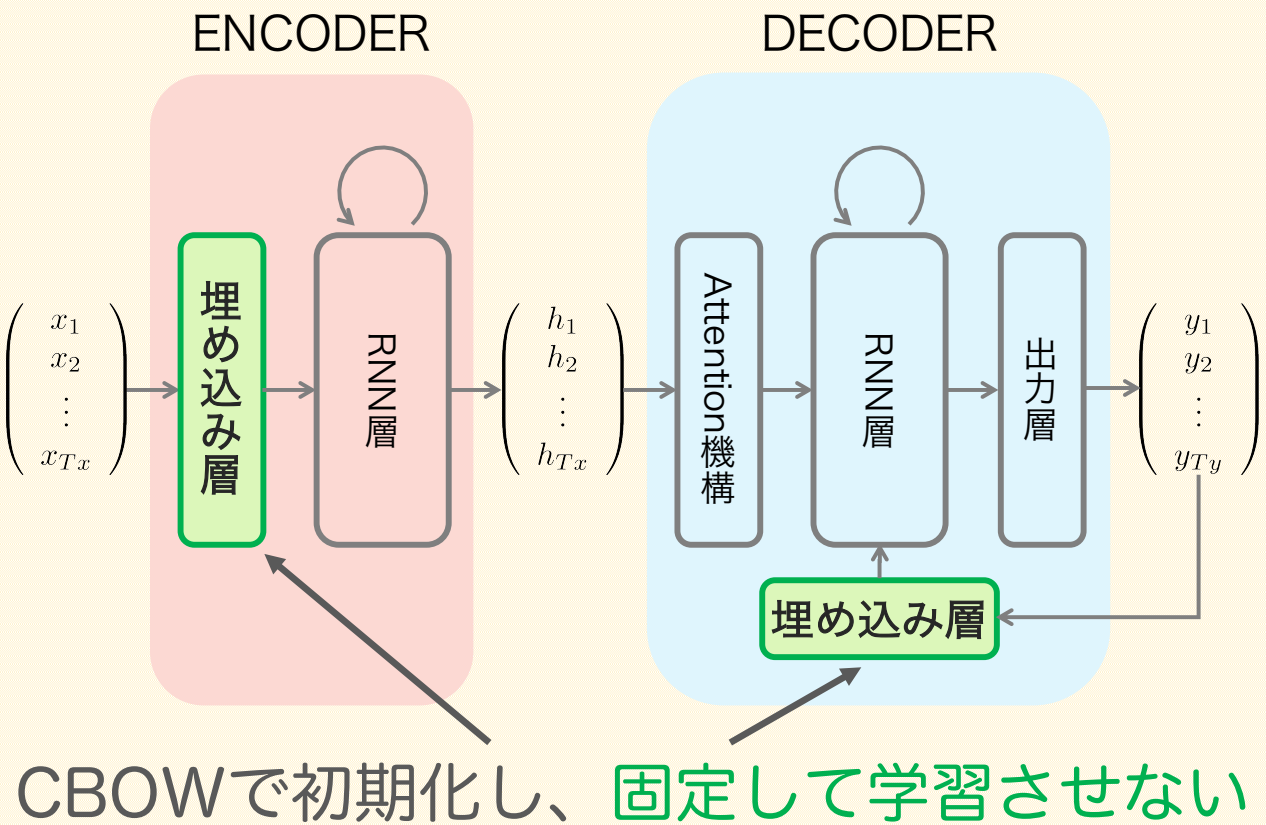
- 2つの埋め込み層のどちらかのみをCBOWで初期化



Encoder側の埋め込み層の方が影響が大きい

初期化後の埋め込み層の固定(1/2)

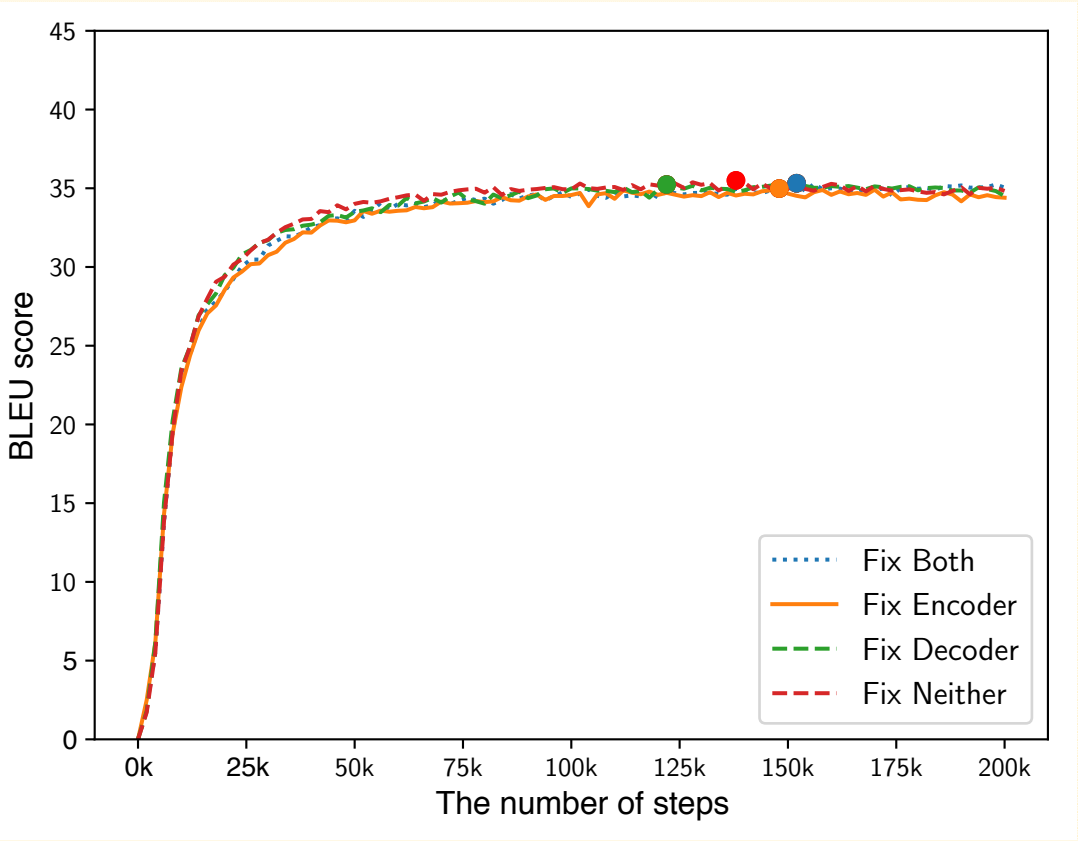
- CBOWで初期化して、固定して学習させずに最終的な値としてそのまま用い、CBOWによる単語埋め込みの翻訳タスクでの有用性を検証



初期化後の埋め込み層の固定(2/2)

- 初期化後の埋め込み層のパラメタを学習させずに固定

Higher is Better



学習曲線はほぼ一致しており、CBOWの単語埋め込みは

NMTモデルの埋め込み層として十分適している。

実験結果

- 埋め込み層の事前学習は、従来のランダム初期化に比べて、翻訳性能の向上と学習の高速化の点で効果的である。
- 対訳データのみでの学習が大規模コーパスを上回った。

- 最適な事前学習手法はCBOW（窓幅5）
- 学習済みのモデルからの埋め込み層の初期化はスコアに影響なし
- Encoder側の埋め込み層の影響力が大きい
- CBOWによる単語埋め込みは翻訳タスクにおいてそのまま使用可能

考察

- CBOWの単語埋め込みは翻訳タスクにおいて適しているため、最適解探索の補助になった
- 異なるドメインを多く含む大規模コーパスよりも、タスクの対訳コーパスのみでの単語埋め込み学習が効果的
- CBOW(5)が最良だったのは、文脈単語の扱い、窓幅共に中間的であることによる
- Encoderの埋め込み層の影響の大きさから、事前学習では入力部に近い方が重要なのでは

まとめ

対訳コーパスを用いてCBOWで事前学習した単語埋め込みでNMTモデルの埋め込み層を初期化

→非常に低コストに翻訳性能の向上、学習の高速化

この手法は埋め込み層を有するあらゆるモデルに適用可能

今後の課題：

異なる対訳コーパスでの検証、異なる構造のニューラル機械翻訳や対話などの機械翻訳以外のタスクへの応用