ニューラル機械翻訳における文脈情報の選択的利用

藤井 諒[1]， 清野 舜[1,2]， 鈴木 潤[1,2]， 乾 健太郎[1,2]

[1]東北大学 [2]理化学研究所AIPセンター

研究背景

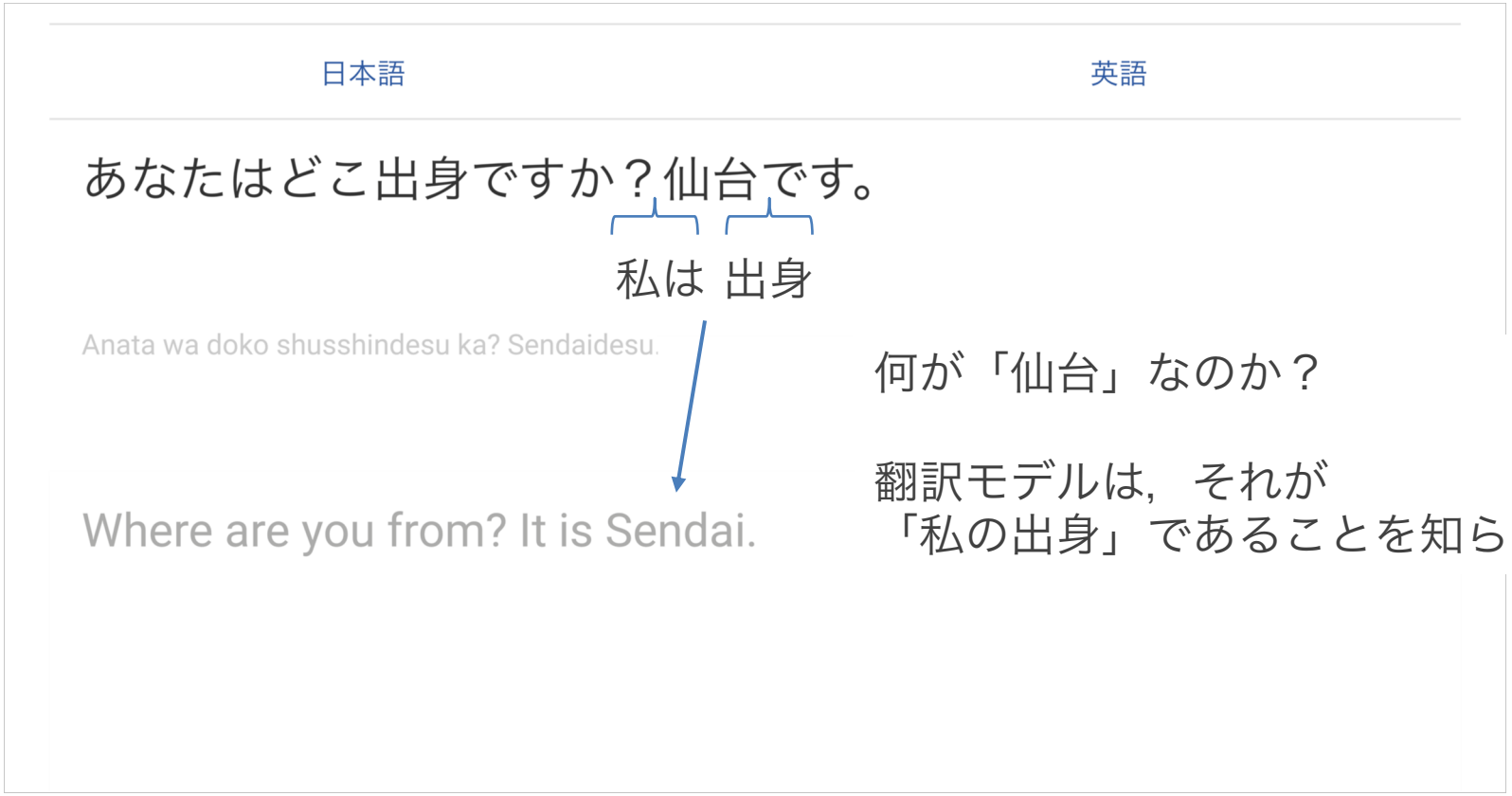
- ニューラル機械翻訳の登場以降，翻訳品質は劇的に向上

単語，フレーズ単位の翻訳精度が向上したことで，

- 代名詞の誤訳，省略の補完，生成文間の語彙一貫性がしばしば問題視
- 翻訳対象文外の文脈利活用が注目されはじめた

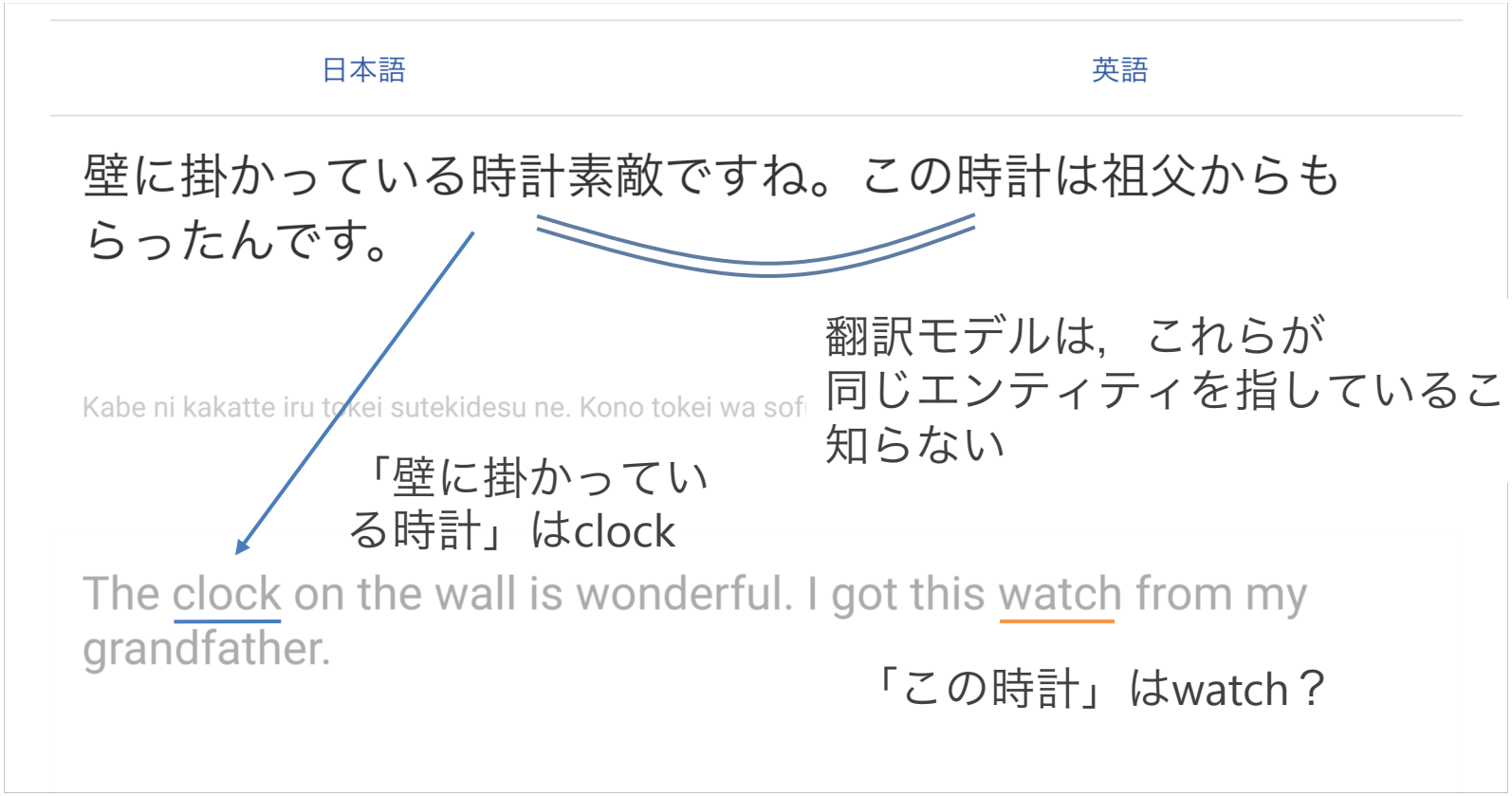
現状の翻訳システムが抱える問題点

- 省略単語の理解ができていない例



現状の翻訳システムが抱える問題点

- 生成文間の語彙一貫性が保たれていない例



機械翻訳における文外文脈への2つのアプローチ

1. モデルを拡張して文外文脈をベクトル化するための機構を追加する方法（Voita+ ’18, Wang+ ’17)

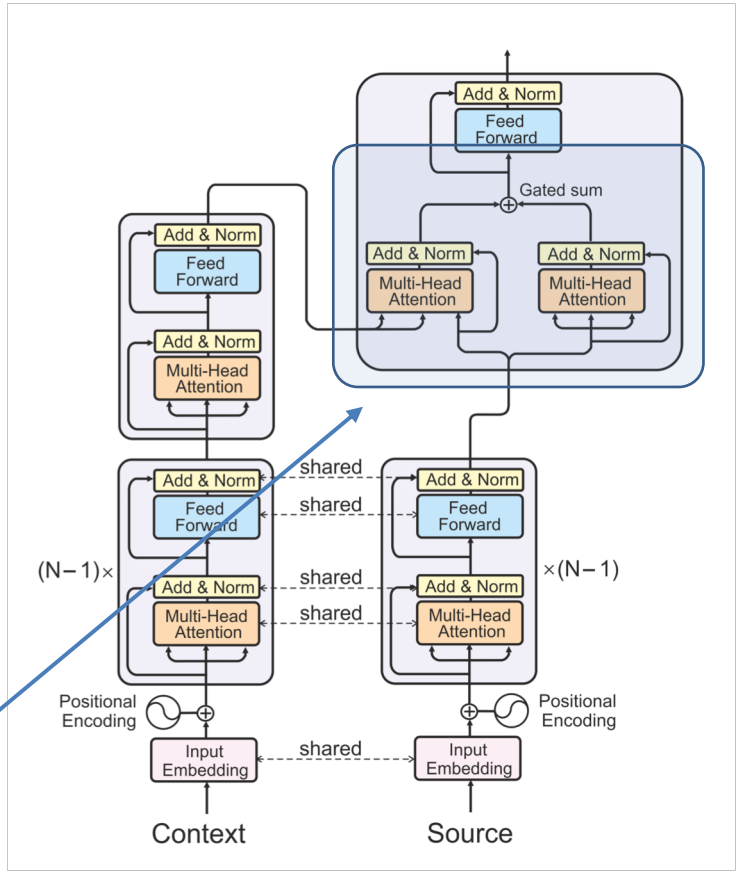
2. データの改変により文脈情報をモデルに同時に扱わせる方法（Tiedemann, Scherrer ‘17）

アプローチ1: モデル側の拡張

モデル側のアプローチからは...

- Transformer（Voitaら），H-LSTM（Wangら）により文外文脈のベクトル表現を作成

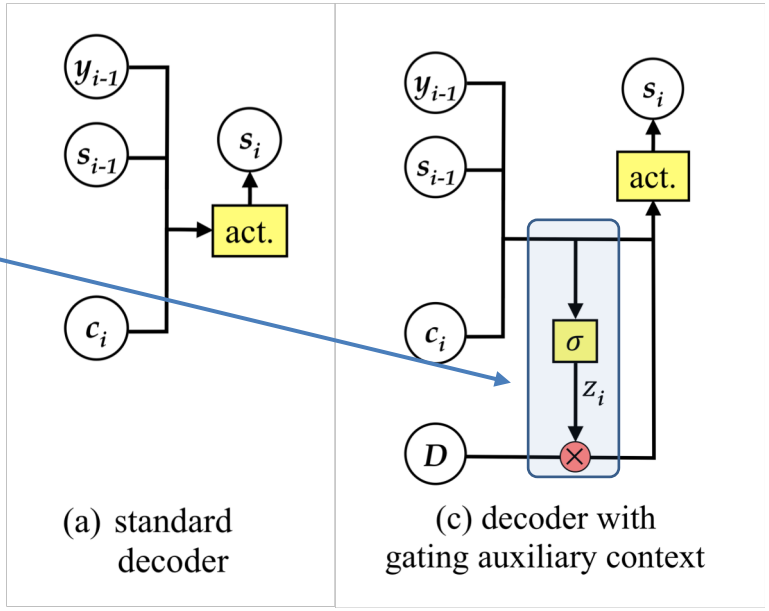
ベクトル表現の混ぜ合わせにはともに**ゲート機構**を用いている



Voita et al. ‘18
Context-Aware Neural Machine Translation Learns Anaphora Resolution

アプローチ1: モデル側の拡張

追加文脈のゲーティングが
翻訳精度の向上に有効
（Wangら）



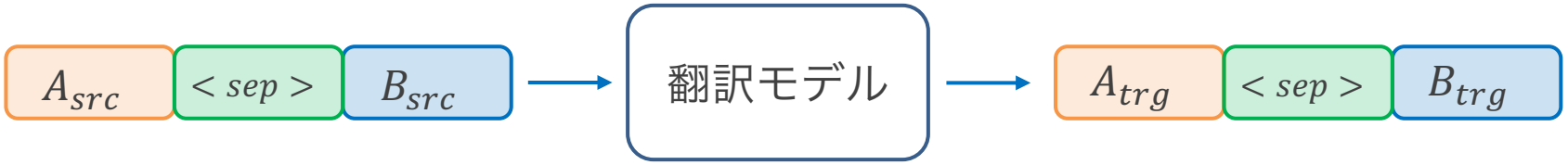
Wang et al. ’17
Exploiting Cross-Sentence Context for Neural Machine Translation

大域的な**文脈情報の必要性は**,
生成単語の曖昧性により**変化**し,
必要に応じた文脈情報の使い分けが必要 -->

アプローチ2: データ側の工夫

データ側のアプローチからは...

- 原言語文，目的言語文の双方あるいは原言語文側のみに前文を結合
- 翻訳単位の拡張により翻訳精度が向上することを指摘



モデルアーキテクチャに依存せず幅広く適用可能 -->

本研究の目的

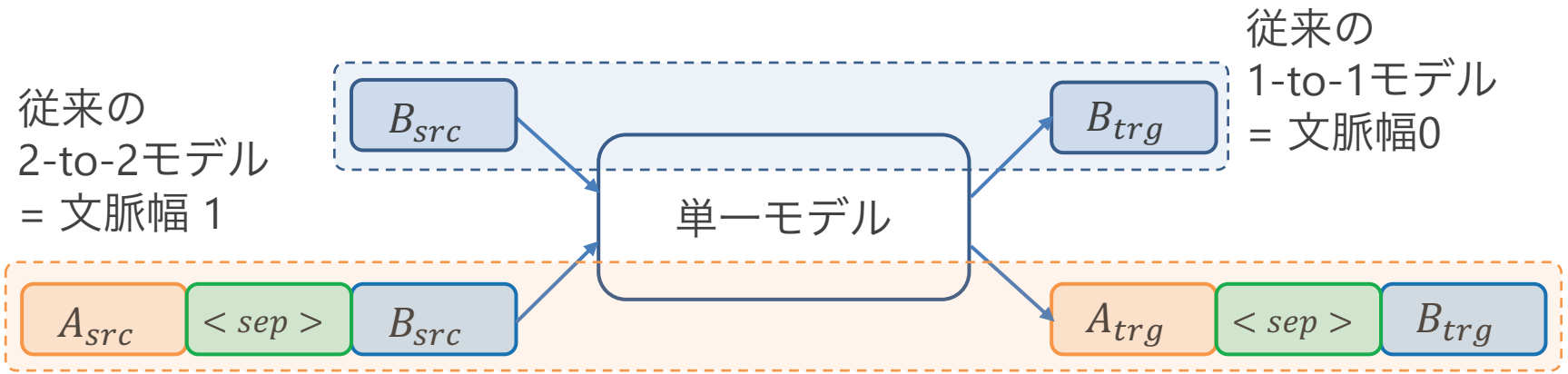
文脈情報をモデルに与えるためには
多様な手法が考えられるが,
どのような情報の加え方が有用なのかは明らかでない

選択的に文脈情報を利用でき,
モデルアーキテクチャ非依存な学習データの与え方
の検討

提案手法

Tiedemann, Scherrerの2-to-2, 2-to-1翻訳を拡張

これらのデータをオリジナルのパラレルデータに加えた,
異なる文脈幅を持つ学習データの混ぜ合わせ学習を提案

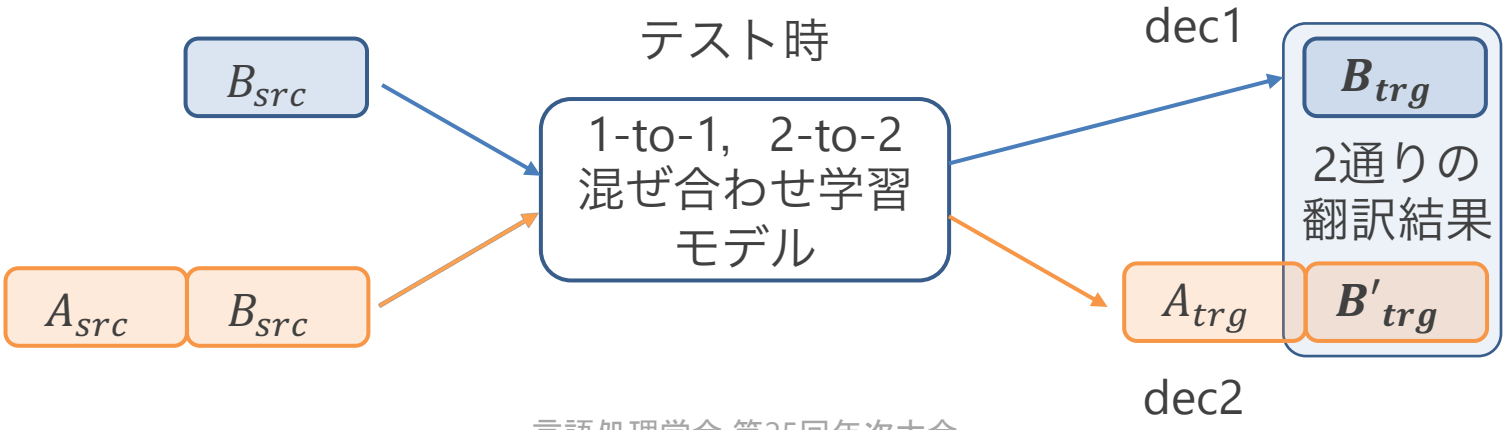


提案手法に対するモチベーション

- 先行文脈が必要である例はコーパス全体に対して多くない

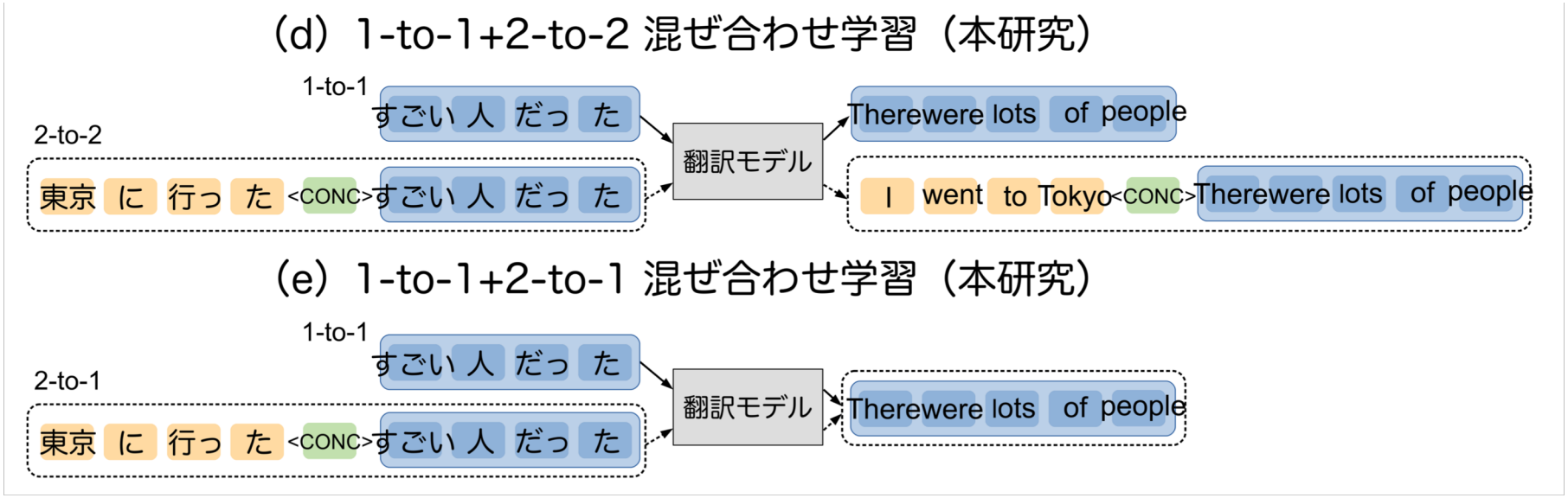
-> 文脈情報を持たない学習データを与えて文単位の対応関係を明らかにすることが有効では？

・単一モデルでありながら2通りのデコード結果を利用可能（一種のマルチタスク学習）



モデルへの入出力

- 入力 -> 単文，または2文を連結した日本語文
- 出力（2-to-2混ぜ合わせ）-> 入力文と同数の英語文
- 出力（2-to-1混ぜ合わせ）-> 後ろの文に対応する英語1文



データセット 1/2

- コーパス

OpenSubtitles2018コーパス

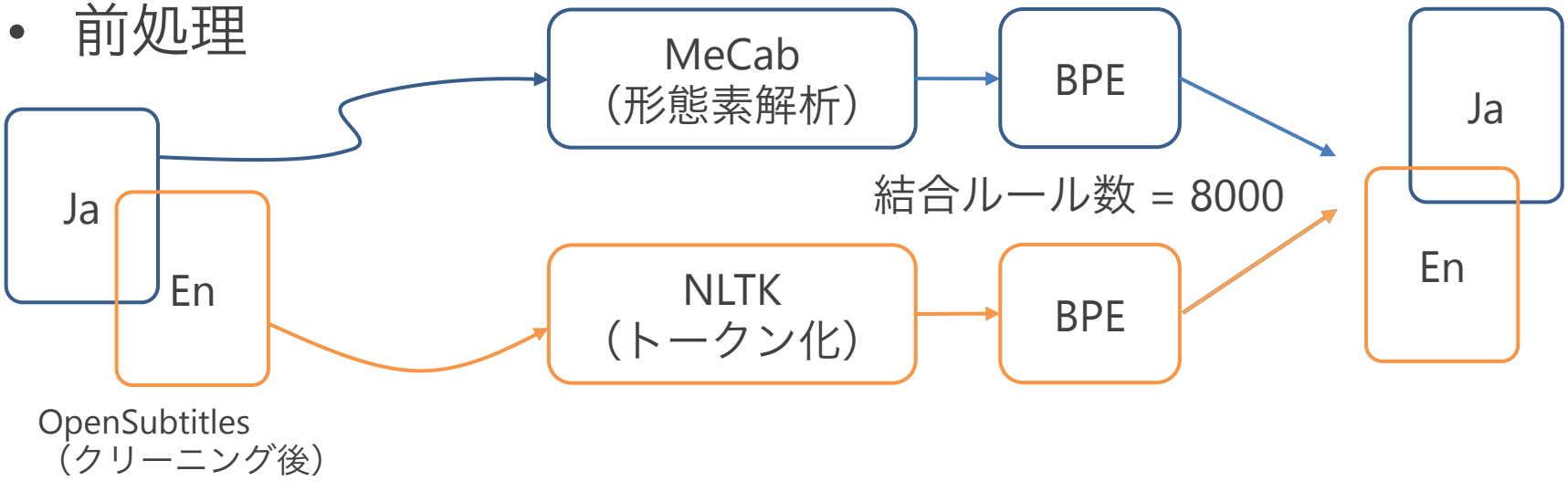
opensubtitles.orgに投稿された映画/ドラマ字幕からなる日英対訳コーパス（210万文）

クリーニングを適用し約半分の110万文を使用

```
% paste OpenSubtitles2018.en-ja.en OpenSubtitles2018.en-ja.ja | head
"THE CABINET OF DR. CALIGARI"   カリガリ博士の小屋
Act 1   幕1
"Spirits surround us on every side - they have driven me from hearth and home, from
wife and child."   私は亡霊のおかげで家族や家庭を捨てざるをえなかった
"She is my bride..."    僕の婚約者だ
"What she and I have experienced is yet more remarkable than the story you have told
 me."      僕と彼女の体験はあなたの物語よりも恐ろしい
I will tell you...      今からそれを話してあげよう
"In the small town, where I was born..."        故郷ハレシュテンバルでの出来事だ
"...a traveling fair had arrived."      お祭りがひらかれた
"Him..."        ある香具師がやってきた
"My friend, Alan..."    親友のアランだ
```

データセット 2/2

- 前処理

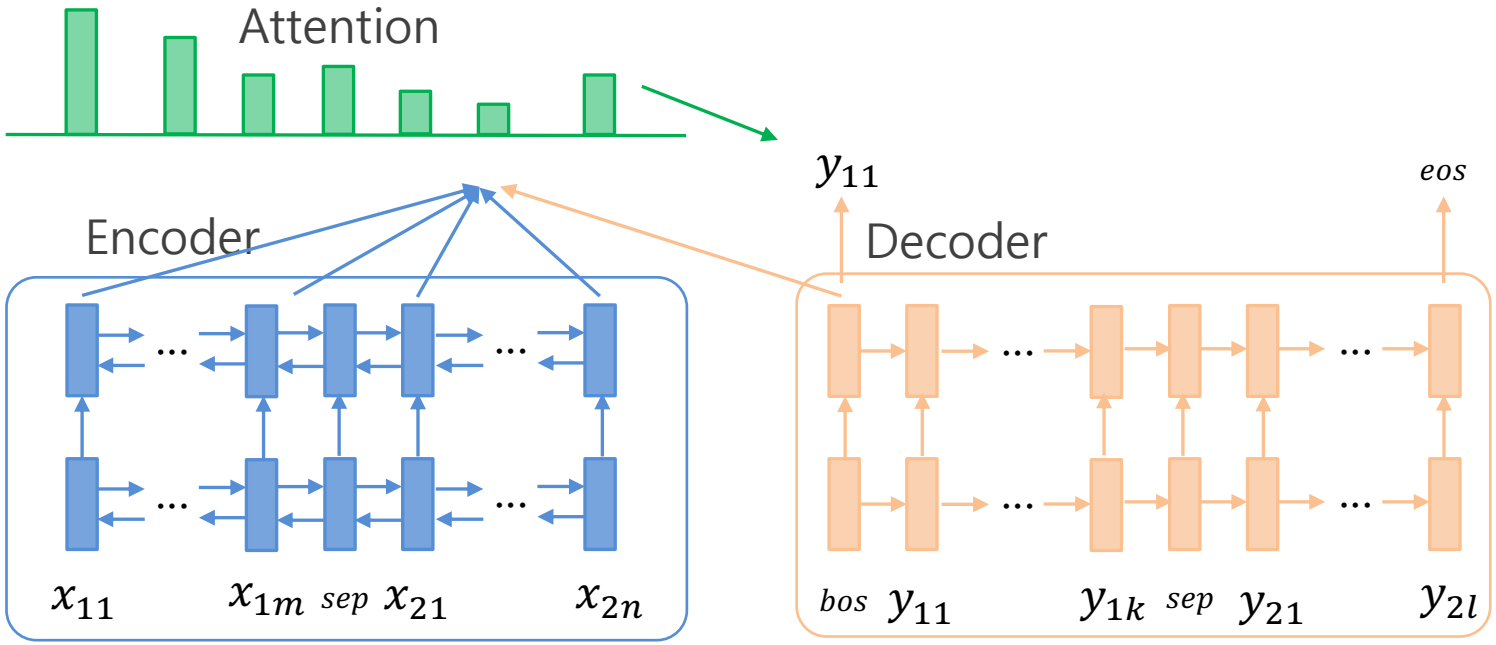


- データの分割
  - 開発用データ: 無作為に選んだ10ストーリー，約6000文ずつ
  - 評価用データ: 同様に作成，JESCのテストデータも使用
  - 学習データ: 残りすべての文対

実験設定

- アーキテクチャ

LSTM（Luongらの設定に準拠，エンコーダはBi-LSTM）
デコードには幅5のビームサーチを適用



実験結果

先行手法と提案法による出力を自動評価指標BLEUにより評価

Model	Train time	Test	OpenSubtitles	JESC
ベースライン	1-to-1	1-to-1	19.84	13.98
Tiedemannら	2-to-2	2-to-2	20.47	15.45
	2-to-1	2-to-1	20.19	14.71
Ours	1-to-1+2-to-2	1-to-1 (dec1)	20.88	17.27
		2-to-2 (dec2)	20.85†	17.35†

最大1.90ポイントの上昇

Table1: 乱数シードを変えて学習した
5モデルのアンサンブルによるスコア
(†：2-to-2に対してp=0.05のブートストラップ検定で有意差)

実験結果

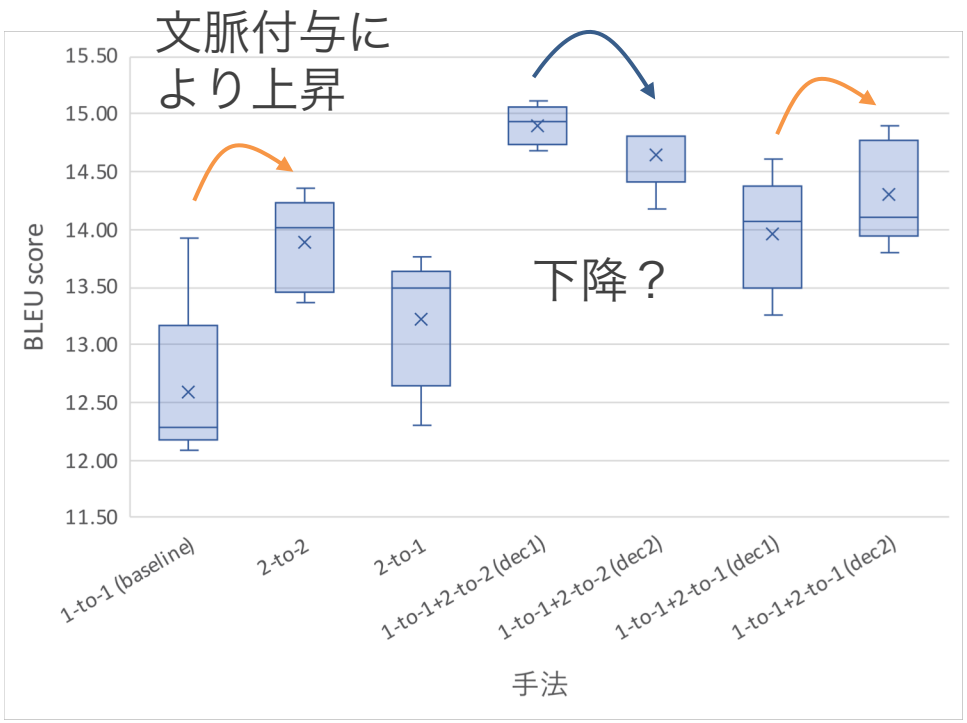


Figure1: 乱数シードに対するBLEUスコアの分布 （JESC）

文脈なしモデルと対応する文脈付与モデルの比較

- 2-to-2混ぜ合わせモデルでは単体モデルの性能はdec2で低下

-> デコード長の影響

- 多くの場合で文脈情報は翻訳精度の向上に寄与

文脈情報の選択的利用に関する検討

混ぜ合わせモデルから得られる2通りのデコード結果から，
常に良い方を選択できる場合のオラクルスコアを調査

Model	Train time	Test	OpenSubtitles	選択された文数
Tiedemannら	2-to-2	2-to-2	20.47	-
Ours	1-to-1+2-to-2	1-to-1 (dec1)	20.88	5264 (83%)
		2-to-2 (dec2)	20.85	1076 **(17%)**
Oracle		文単位BLEUの高い方を選択	21.96	6340

BLEUスコアのオラクルはdec1，またはdec2のみを用いる場合からさらに1.08ポイント上昇

文脈情報の選択的利用に関する検討

- 文脈必要性の判断による**事前選択**

それぞれのインスタンスに対し，文脈情報を与えるべきかを判断できる分類器に事前に通す方法

- 2通りの出力後**事後選択**

事前学習済み単語ベクトルなどを用いて文ベクトルの類似度に基づく候補選択

-> 具体的な選択手法の考案は今後の課題

事例研究

（a）

先行文脈 : 倉庫に戻り 違う道を探した方がいいかも。 他の**通路がない。**

翻訳対象文 : これしか。

参照訳 : there is no other way .

ベースライン， 提案手法dec1 : this is it .

2-to-2， およびdec2 : this is the only **way** .

前文を参照できるおかげで、「これしか」ないものが通路であることがわかった

事例研究

（b）

先行文脈 : アンデゥルー・ブラウナーの退職の ファクスを見て驚きました。退職の手紙？

翻訳対象文 : そう，電話もつながらない。

参照訳 : a resignation ? yeah, i ca n’t get him on the phone .

2-to-2 : a retirement letter .

提案手法dec2 : yeah, i did n’t call him .

2-to-2では文単位の
対応関係を誤っている
1-to-1データによる
明示的な文アライメント
が有効？

まとめ

- モデルに対する文脈情報の与え方の検討として先行手法に加え，**異なる文脈幅を持つ学習データの混ぜ合わせ学習**を提案

- 映画字幕翻訳タスクにおいて既存手法の出力に匹敵，あるいは上回る翻訳精度を達成

- 提案モデルが，学習データの与え方により従来の1–to-1, 2-to-2の2つのモデルからの出力に類似した出力を使い分けられる可能性を確認（実際に選択手法についての検討を行うのは今後の課題）