スタイル変換技術による対訳コーパスから同時通訳コーパスへの拡張

2019/08/30

奈良先端科学技術大学院大学
二又航介、須藤克仁、中村哲

背景：同時通訳 (Simultaneous interpretation)

原言語の入力文の終了を待たずに目的言語への訳出を開始

- 通常の翻訳システムでは原言語の入力終了を待つため遅延が発生
- 講義や講演など主に訳出の遅延が許されない場面で使用
- 遅延を最小限にしつつ正確に部分訳出を行う

同時通訳による訳出例

A brand-new computer on the desk / which my father gave me on my birthday / doesn't work now.

原言語文

机上にある新しいコンピュータですね、

これは父から誕生日にもらったものです、

ですが、今故障しています。

目的言語文

背景：通訳方法の違いによる遅延

英日翻訳のように語順が大きく異なる言語間の翻訳では訳出開始までに遅延が発生

- 主要部先行型言語(head-initial)と主要部後続型言語（head-final)の違いによる遅延
- 長い修飾部を持つ英文の訳出開始までに大きな遅延が発生
- 原文の節や句の順序を守りながら原言語の語順に近い形で訳出（順送り）することで遅延が少なくなる[1]
- 順送り方式による訳出では語順の洗練性は無いが助詞等による致命的な間違いが発生しない

[1] 水野的：同時通訳の理論 —認知的制約と訳出方略，朝日出版社 (2015).

背景：通訳方法の違いによる遅延

訳出方法の違いによる遅延

A brand-new computer on the desk which **my father gave me** on my birthday doesn't work now .

父から 誕生日 に 貰った 、 机上 に ある 新しい コンピュータ は 今 故障 しています 。

(待ち時間...) 父から 誕生日 に 貰った 、 机上 に ある 新しい コンピュータ は . . .

訳出開始までに大きな遅延が発生する例

A brand-new computer **on the desk** which my father gave me on my birthday doesn't work now .

机上 にある 新しい コンピュータ ですね、これは 父 から 誕生日 に 貰った もの です 、ですが 今 故障 しています 。

(待ち時間...) **机上 にある** 新しい コンピュータ ですね 、これは 父 から 誕生日 に 貰った もの です . . .

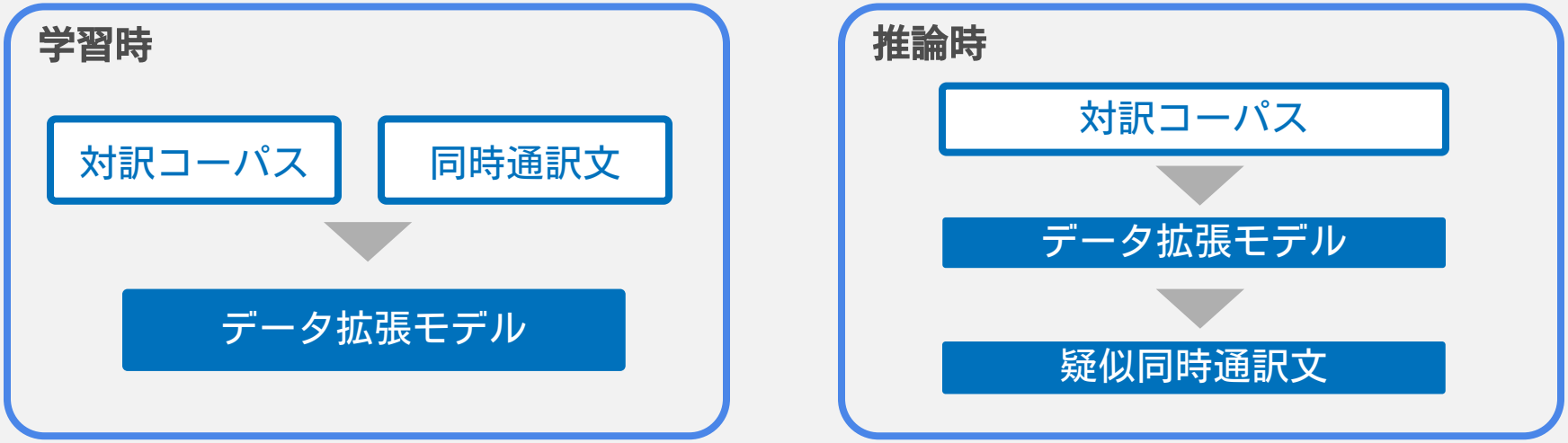
訳出開始までの遅延が少ない例(順送り方式)

目的：対訳コーパスから同時通訳コーパスの拡張

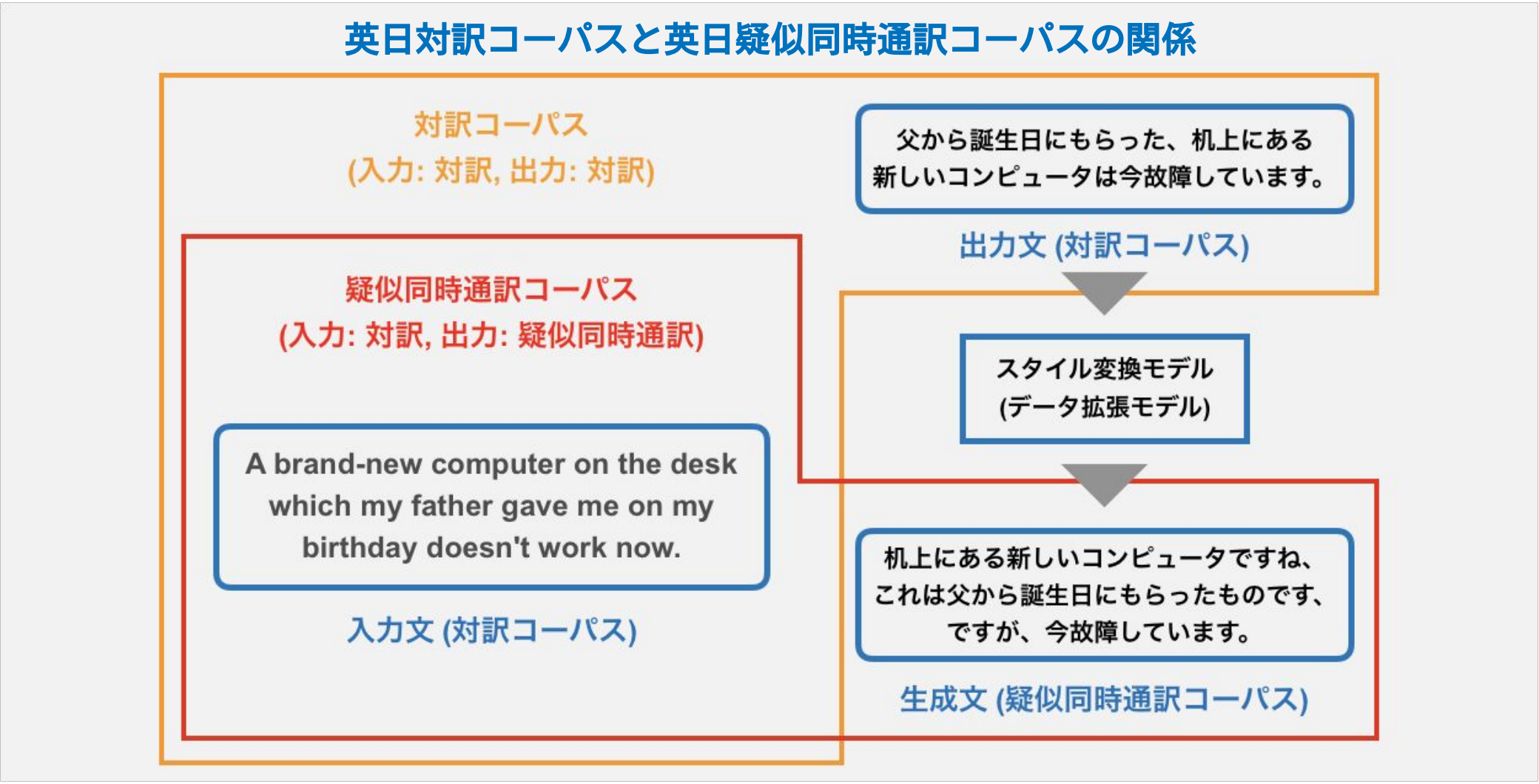
スタイル変換技術によって対訳コーパスから順送り方式の同時通訳コーパスを作成

- 順送り方式で翻訳されたコーパスは少数
- 大量の対訳コーパスは利用可能
- 対訳コーパスの目的言語を順送りの同時通訳文にスタイル変換

対訳コーパスから同時通訳文の拡張過程



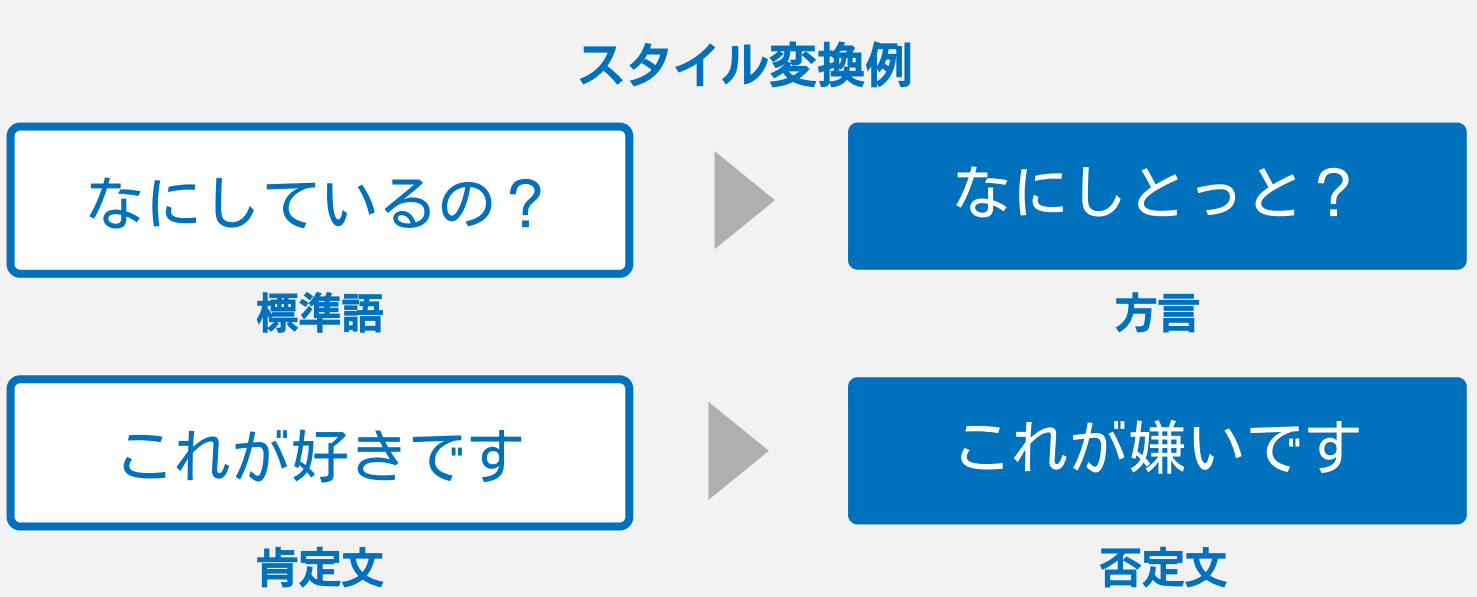
目的：擬似英日同時通訳コーパスの作成



関連研究：スタイル変換

入力文における文意や意図を変更することなく文体(スタイル)を自動的に制御するタスク[2]

- 入力文を意味的に等価な文へ書き換える言い換え生成の一種
- スタイル変換前後の対訳ペアを必要としない
- 変更するスタイル情報が明確な場合に特に有効

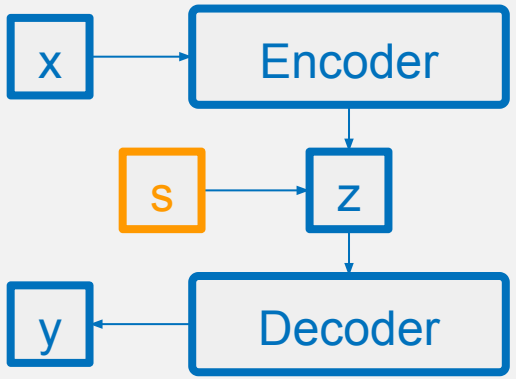


関連研究：スタイル変換の手法

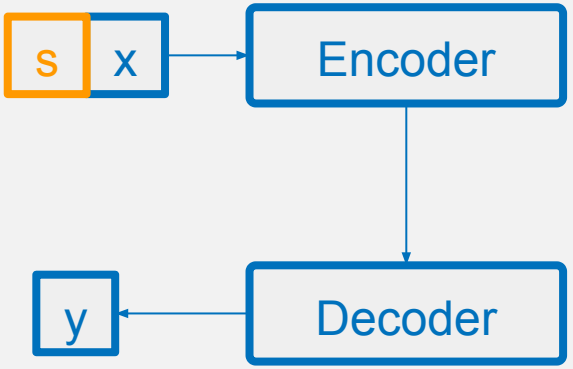
1. **文意とスタイルを分離する手法**[3]
   - スタイル変換の一般的な手法
   - スタイル情報と独立した文意の潜在表現を学習
2. **文意とスタイルを分離せず直接スタイル変換する手法**[4]
   - 潜在表現には文意とスタイル情報が含まれる
   - あるスタイルから異なるスタイルへ直接変換

x: 文章(スタイル1)
y: 文章(スタイル2)
s: スタイル識別子
z: 潜在表現

2種類のスタイル変換手法



文意とスタイルを分離する手法



直接スタイル変換する手法

[3] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, Rui Yan, Style Transfer in Text: Exploration and Evaluation, 2017
[4] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, Y-Lan Boureau, Multiple-Attribute Text Rewriting, 2019

関連研究：Style Transformerによるスタイル変換

Style Transformer[5]

- Transformerをベース
- 直接スタイルを変換
- 意味的な単語を変換する

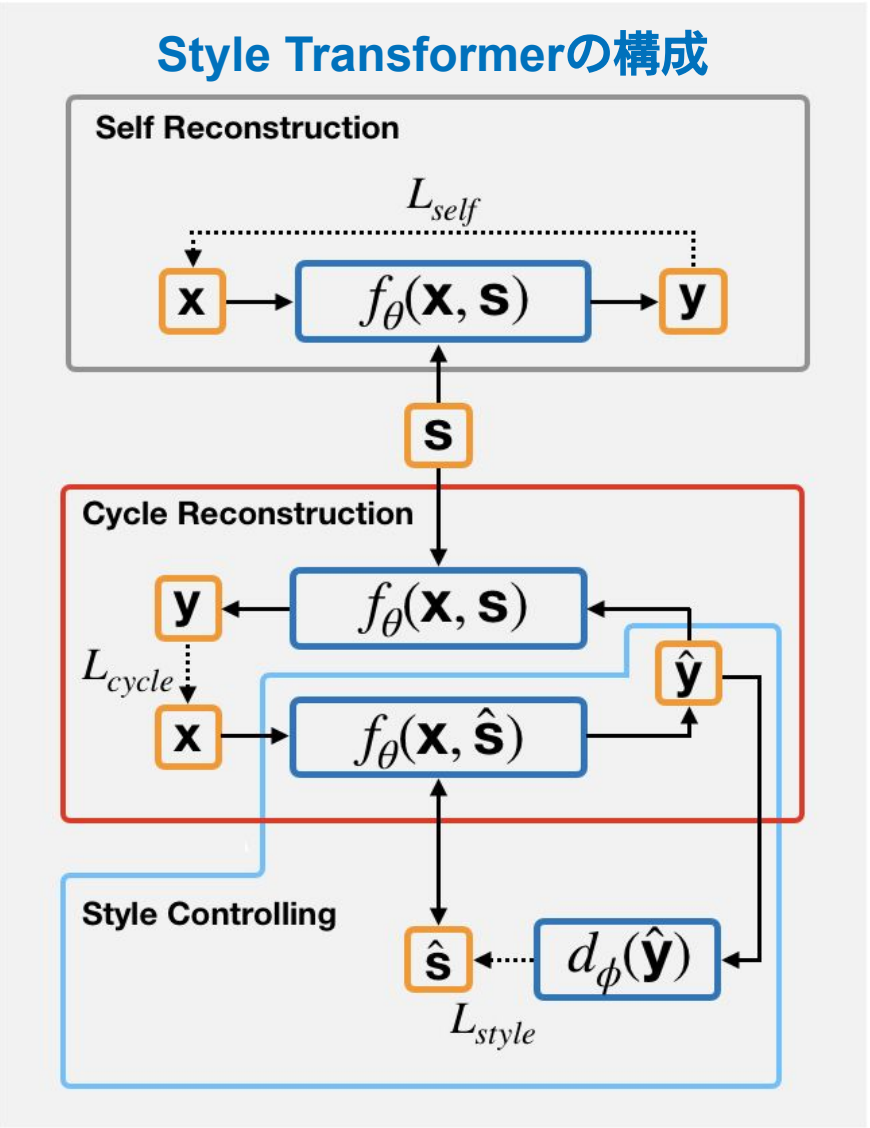
x: 入力文(スタイル1)

y: 変換文(スタイル1)

ŷ: 変換文(スタイル2)

s: 変換前のスタイル識別子

ŝ: 変換後のスタイル識別子



[5] Ning Dai, Jianze Liang, Xipeng Qiu, Xuanjing Huang, Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation, 2019

関連研究：Style Transformerによるスタイル変換

Self Reconstruction

入力文(**x**)とスタイル(**s**)は同じ

$$\mathcal{L}_{self}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|\mathbf{x}, \mathbf{s})$$

Cycle Reconstruction

入力文(**x**)とスタイル($\hat{\mathbf{s}}$)が異なる

入力文を異なるスタイル($\hat{\mathbf{s}}$)で変換後($\hat{\mathbf{y}}$)
スタイル(**s**)によって入力文を復元(**y**)

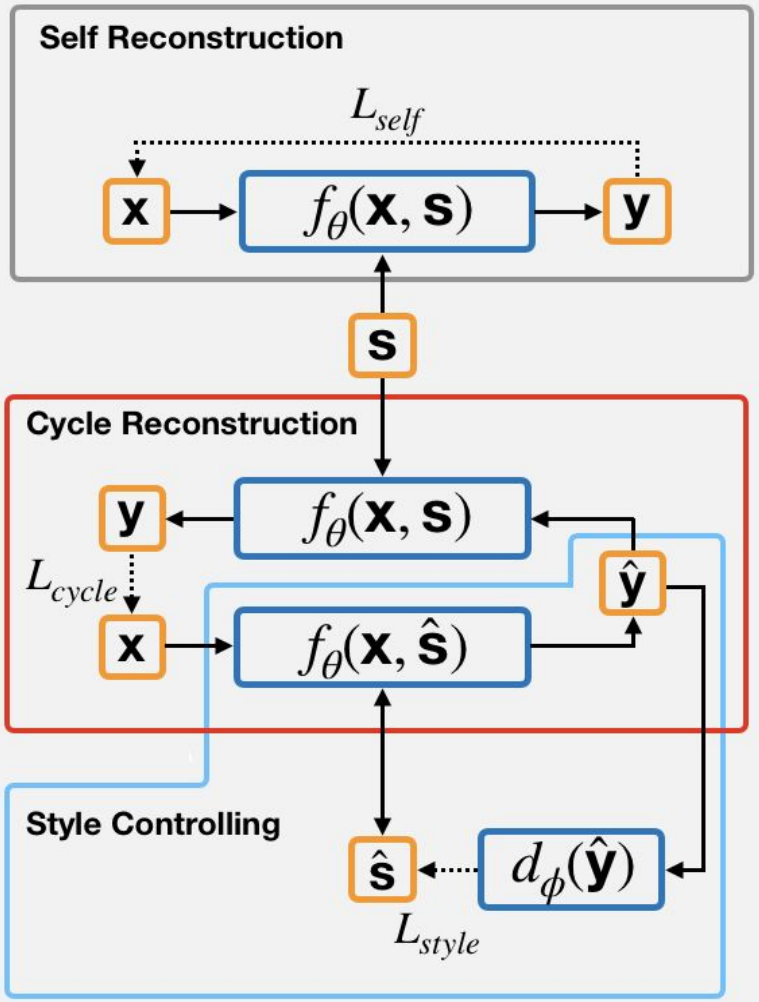
$$\mathcal{L}_{cycle}(\theta) = -p_\theta(\mathbf{y} = \mathbf{x}|f_\theta(\mathbf{x}, \hat{\mathbf{s}}), \mathbf{s})$$

Style Controlling

Cycle Reconstructionにより$\hat{\mathbf{y}}$の
文意が大きく変わるのを防ぐ

$$\mathcal{L}_{style} = -p(\mathbf{c} = \hat{\mathbf{s}}|f_\theta(\mathbf{x}, \hat{\mathbf{s}}))$$

Style Transformerの構成

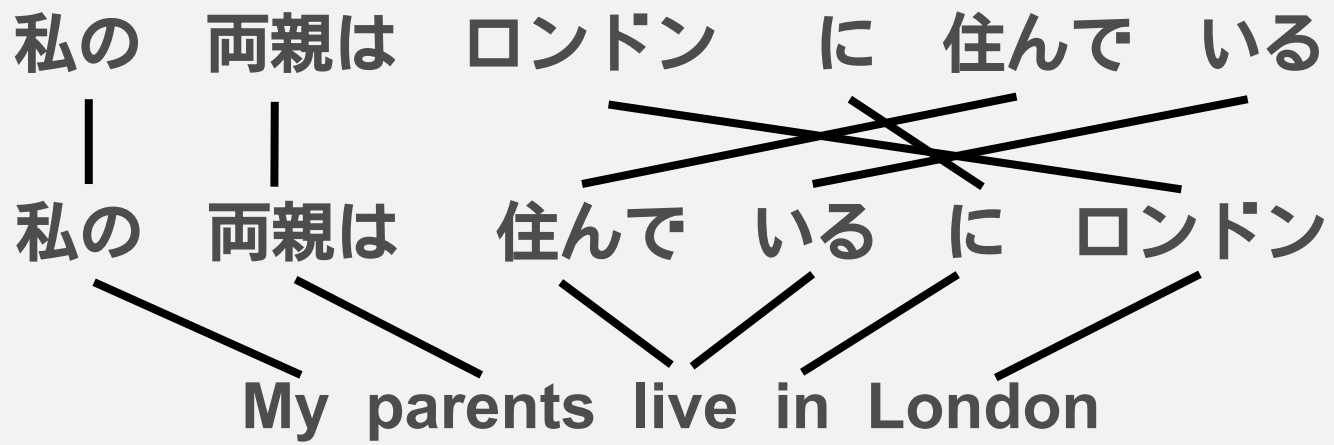


関連研究：事前並べ替え[6]

翻訳機に入力する前に原言語における文の語順を目的言語の語順に近づくように並び替える手法

- 主に統計的機械翻訳(SMT)で使用
- 順送り方式の同時通訳文は原言語の語順と類似している
- 事前並べ替えによって原言語との単語間の交差が少なくなる

日本語文に対する事前並べ替えの適用例



[6] Tetsuji Nakagawa, 2015, Efficient Top-Down BTG Parsing for Machine Translation Preordering, in Proceedings of ACL, pages 208-218.

提案手法：スタイル変換と事前並べ替えによるデータ拡張

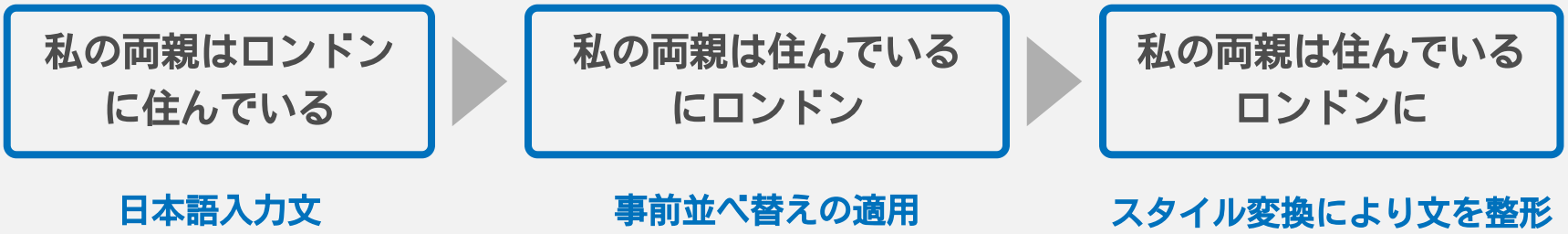
スタイル変換前の対訳コーパスに事前並び替えを適用

- スタイル変換は単語の置換や単語の削除など表現力が限定的であり語順の並べ替えには不向き
- スタイル変換前に語順を並べ替えることでスタイル変換を容易に

スタイル変換を用いたデータ拡張の過程

1. 日本語入力文を英文の語順のように事前並べ替え
2. スタイル変換により文を正しく整形

提案手法による同時通訳コーパスの拡張方法



実験設定：使用データ

Style Transformerと事前並び替えによって英日対訳コーパスから英日同時通訳コーパスへのスタイルを変換

- 事前並べ替えの適用あり/なしの2種類で実験
- 対訳コーパス：日本語話し言葉コーパス(CSJ)
- 同時通訳コーパス：TEDコーパス(独自に収集した同時通訳文)
- 事前並べ替え：ASPEC

CSJ,TED,ASPECのコーパスサイズ

Corpus	Number of sentences		
	Train	Val	Test
CSJ	3509	500	500
TED	28606	500	500
ASPEC	1000000	1790	1812

スタイル変換を行うコーパスペア

Corpus pair	Detail
CSJ-TED	CSJ から TED
CSJ(preordered)-TED	事前並べ替えした CSJ から TED
CSJ-CSJ(preordered)	CSJ から事前並べ替えした CSJ

実験設定：自動評価

スタイル変換で主に使用される3つの評価指標

Style accuracy

- 二値分類器（対訳or同時通訳）によってスタイル分類精度を計測
- スタイル変換された文（疑似同時通訳文）が同時通訳のスタイルとして識別されることを期待
- CSJとTED(同時通訳文)により学習

Bleu score

- Bleu scoreによってスタイル変換後の文意の保持具合を計測
- スタイル変換された文（疑似同時通訳文）が入力文と同一の文意を保持していることを期待

Perplexity

- Perplexityによってスタイル変換後の文章の流暢さを計測
- スタイル変換された文（疑似同時通訳文）が入力文と同様に流暢であることを期待

自動評価による実験結果

CSJ-TED: 語尾や単語を変更する傾向

- 文意と流暢さが保持されたため高BLEU, 低Perplexity

CSJ(preordered)-TED: 語順変化&単語追加する傾向

- 語順変化&単語追加されたため低BLEU, 高Perplexity

CSJ-CSJ(preordered): 長い文を短い単位に区切る傾向

- 語順変化&流暢さが損なわれたため低BLEU, 高Perplexity

自動評価指標による実験結果

Corpus pair	ACC	BLEU	PPL
CSJ-TED	89.0	48.8	73.5
CSJ(pre-ordered)-TED	96.0	21.8	361.5
CSJ-CSJ(preordered)	30.2	28.3	242.7

実験結果: CSJ-TEDの生成例

CSJからTEDへのスタイル変換

- 語尾や単語を同時通訳コーパスに現れるものに変換する傾向
- Style accuracy: 89.0, Bleu score: 48.8, Perplexity: 73.5

入力文(CSJ)	世界 戦争 が ヨーロッパ から 始まり ました 。
変換文(TED)	世界 戦争 が ヨーロッパ から 始まり ます 。

入力文(CSJ)	私 が 言わ なくちゃ いけ ない 内容 に 入る 前 に , 少し 私 の 自己 紹介 です 。
変換文(TED)	私 が 言わ なくちゃ いけ ない 内容 に 入る 前 に , 少し 僕 の 自己 紹介 です 。

入力文(CSJ)	そして , この 町 の 中 で 最悪 の 裁判 所 でし た 。
変換文(TED)	そして , この アフリカ の 中 で 最悪 の 裁判 所 でし た 。

実験結果: CSJ(preordered)-TEDの生成例

CSJ(preordered)からTEDへのスタイル変換

- 語順変更や単語追加が行われるが文意が大きく変わる傾向
- Style accuracy: 96.0, Bleu score: 21.8, Perplexity: 73.5

原文(CSJ)	学習 データ は 、こちら と 同じ もの です 。
入力文(CSJ(preordered))	学習 データ は 、です 同じ もの と こちら 。
変換文(TED)	学習 データ は ない です 、同じ もの と こちら 。

原文(CSJ)	一方 アジア 人 は 、日本人 学 部 生 は 、僕 一人 で 少ない 方 でし た 。
入力文(CSJ(preordered))	た 一方 アジア 人 は 、日本人 学 部 生 は 、でし 僕 一人 で 方 少ない 。
変換文(TED)	私 一方 アジア 人 は 、日本人 学 部 生 、一人 で 少ない 。

実験結果: CSJ-CSJ(preordered)の生成例

CSJからCSJ(preordered)へのスタイル変換

- 長い文を短い単位に区切り同時通訳らしい文を生成する傾向
- Style accuracy: 30.2, Bleu score: 28.3, Perplexity: 242.7

入力文(CSJ)	講演 音声 認識 の 識別 率 は 、今 の ところ 七十 パーセント 程度 です 。
参照文(CSJ)	識別 率 の 講演 音声 認識 は 、です 程度 ところ の 今 七十 パーセント 。
変換文 (CSJ(preordered))	講演 音声 認識 の 識別 率 は です 今 の ところ 七十 パーセント 程度 。

入力文(CSJ)	次 に 本 研究 に おけ ます システム の概要 を 説明 いたし ます 。
参照文(CSJ)	ます いたし 次 に 概要 の おけ ます に 本 研究 システム を 説明 。
変換文 (CSJ(preordered))	つい に 本 研究 に おけます システム の 概要 を きちんと 説明 て 。

実験結果: CSJ-CSJ(preordered)の生成例

入力文(CSJ)	使用 する 関係 として は 、動詞 目的 語 名詞 という この 三 種類 を 使用 します 。
参照文(CSJ)	ます し は 、を 使用 この 三 種類 いう と 動詞 目的 語 名詞 し と 関係 する 使用 て 。
変換文 (CSJ(preordered))	使用 する 関係 とし ます 動詞 目的 語 名詞 という この 三 種類 を 使用 して 。

入力文(CSJ)	これ は 、係り受け の 係り 先 が 間違っ て おり ます の で 誤り と なっ て おり ます 。
参照文(CSJ)	ます おり なっ と 誤り て で の ます これ おり 先 係り の は 、係り受け が っ 間違 て 。
変換文 (CSJ(preordered))	これは 、係り受け です 係り 先 が間違 っ て です おり ま す の で 誤り と なり 。

入力文(CSJ)	次 に 実際 に 決定 木 を 構築 し て の 選択 し 検索 を 行なう 実験 という の を 行ない ました 。
参照文(CSJ)	たし 選択 の し 次 に 実際 決定 木 を 構築 に て まし 行ない の いう と 実験 行なう 検索 を を 。
変換文 (CSJ(preordered))	次 に 実際 に 決定 木 を 構築 し ます 選択 し 検索 を 行ない ます 実験 という の を 行ない ました 。

実験設定：人手評価

既存の自動評価指標で疑似同時通訳文を評価するのは困難

- 長い文を短く区切る傾向にあり最も同時通訳文らしい CSJ-CSJ(preordered)のペアを対象に評価実験
- 7人の被験者が50サンプルを評価
- スタイル変換前の文（入力文）とスタイル変換後の文（生成文）を提示
- 評価指標に当てはまるか1から5の5段階で評価

評価指標

- 入力文と比較し生成文が短い単位に区切られているか(Segmentation)
- 生成文は日本語として自然で流暢であるか(Fluency)
- 入力文と生成文が意味的に同一であるかどうか(Identity)

人手評価による実験結果

SegmentationとFluencyの間, SegmentationとIdentityの間にトレードオフの関係性

- Segmentation ≥ 3.0，Fluency ≥ 3.0，Identity ≥ 3.0はそれぞれの平均値が3.0以上であったサンプル
- Segmentationの値が上がるとFluencyとIdentityの値が下がる
- Segmentationの値が下がるとFluencyとIdentityの値が上がる

7人の被験者による評価平均値

	Num samples	Segmentation	Fluency	Identity
All samples	50	3.03	2.42	3.11
Segmentation $\geq$ 3.0	30	3.72	2.02	2.95
Fluency $\geq$ 3.0	10	2.37	4.00	3.74
Identity $\geq$ 3.0	28	2.96	2.76	3.73

人手評価による実験結果

Segmentationの平均値が特に高くFluencyとIdentityの平均値が特に低い変換例

- 入力文を短い単位に区切る傾向にあるが日本語が不自然

入力文(CSJ)	第 四 声 は 、発音 が 高い ところ から 最も 低い ところ へ 急激 に 移行 し て いき ます 。
参照文(CSJ)	ます いき し 第 四 声 に 急激 移行 へ ところ は 最も 低い から ところ 、発音 が 高い て 。
変換文 (CSJ(preordered))	四 声 は 、 発音 が 高い ます 、 最も 低い 事故 へ 移行 し て 。

入力文(CSJ)	大体 この くらい で 終わり に したい と 思い ます 。
参照文(CSJ)	ます 大体 に 終わり したい と 思い で この くらい 。
変換文 (CSJ(preordered))	ます 、思い くらい で 、終わり に したい と 思い よく 。

人手評価による実験結果

Segmentationの平均値が特に低くFluencyとIdentityの平均値が特に高い変換例

- 入力文が短い単位に区切られないが日本語として自然

入力文(CSJ)	特に です ね 自転車 を 使っ た 旅 が 好き です 。
参照文(CSJ)	です 特に 旅 を 使っ ね です 自転車 た が 好き 。
変換文 (CSJ(preordered))	特に です ね 自転車 を 使っ た 旅 が 好き 。

入力文(CSJ)	で この 町 の 中心 で ある 公園 に ついて 御 説明 し たい と 思い ます 。
参照文(CSJ)	で ます 思い と たい し で ついて 御 説明 に 公園 ある 中心 の この 町 。
変換文 (CSJ(preordered))	で この 町 の 中心 で ある 公園 に ついて 御 説明 し たい と 思い ます 。

まとめ

スタイル変換手法によって対訳コーパスから同時通訳コーパスへと拡張する手法を提案

- スタイル変換と事前並べ替えにより擬似同時通訳文を生成
- 提案手法により長い入力文を短い単位に区切ることが可能
- 提案手法ではSegmentationとFluency, Identityの間にトレードオフ

同時通訳文の特性を考慮したモデル及び指標の導入

- 実際の同時通訳文には語順変化や単語の追加など様々な要因が関与
- 流暢性、意味的同一性を担保した疑似同時通訳文の生成が必要
- 疑似同時通訳文を評価するための新たな指標の導入