P5-28　英日同時通訳システムのための擬似同時通訳コーパス自動生成手法の提案

奈良先端科学技術大学院大学　二又航介　須藤克仁　中村哲

❶ 背景

同時通訳では入力文の終了を待たずに目的言語へ訳出を開始するため遅延が少ない

- 講演など訳出遅延が許されない場面で使用
- 英日翻訳のように語順が異なる言語間の翻訳では訳出開始までに大きな遅延が発生

順送り方式による英日翻訳

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することで遅延が減少[1]

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

日本語訳文: **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。

訳出タイミング: (待ち時間 . . .) 父から誕生日に貰った，机上にある新しい . . .

訳出開始までに大きな遅延が発生する例

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

日本語訳文: **机上にある**新しいコンピュータですね、これは父から貰ったものです、ですが今故障しています。

訳出タイミング: (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)

図1. 訳出方法の違いによる遅延

- 順送り方式の同時通訳コーパスは少数
- 英日対訳コーパスは多数利用可能
- 英日対訳コーパスから順送り方式の英日疑似同時通訳コーパスを作成

同時通訳コーパスにより同時通訳システムを学習させることで翻訳時の遅延減少が期待

❷ 提案手法

文節単位の事前並べ替えとスタイル変換によって擬似同時通訳コーパスを作成

1. 事前並べ替え：順送り文らしい語順に変換
2. スタイル変換：自然で流暢な文に整形

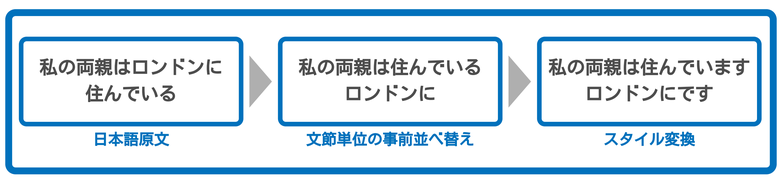

図2. 擬似同時通訳コーパスの生成過程

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

文節単位の事前並べ替え

原言語文の語順を目的言語の語順に類似するように並べ替える手法[2]

- 事前並べ替えにより単語間交差が減少
- 順送りの同時通訳文は原言語の語順と類似
- 文節単位の事前並べ替えにより単語単位より同時通訳文らしい自然な語順に変換

表1. 文節単位の事前並べ替え適用例

英語原文	The proposal technique was verified by the mounting example.
日本語原文	提案 手法 を 実装 例 に より 検証 した 。
並べ替え文(単語)	提案 手法 を し た 検証 に より 実装 例 。
並べ替え文(文節)	提案 手法 を 検証 した 実装 例 に より 。

❸ 実験

提案手法により作成した擬似同時通訳コーパスの品質を自動評価及び人手評価により評価

- スタイル変換モデルには教師なしニューラル機械翻訳を使用[3]

表2. スタイル変換に使用した各コーパスの統計情報

Corpus	Detail	Train	Val	Test
ASPEC(JA)	対訳コーパスにおける日本語文	1,000,000	1790	1812
ASPEC(preordered)	ASPEC(JA)に事前並べ替えを適用した日本語文	1,000,000	1790	1812
TED(SI) + CSJ + OS + JESC	同時通訳文TED(SI)にCSJ, OS(OpenSubtitles), JESCを加えた口語文	810,134	-	-

表3. スタイル変換を適用する方向

Corpus Pair	Detail
ASPEC(JA) → ASPEC(preordered)	対訳コーパスから事前並べ替え文への変換
ASPEC(preordered) → TED(SI) + CSJ + OS + JESC	事前並べ替え文から口語文への変換

自動評価実験

スタイル変換モデルの自動評価指標により評価

表4. 擬似同時通訳文の評価に用いた自動評価指標

Metrics	Detail
Style Accuracy(ACC)	2値分類器により擬似同時通訳文のスタイル分類精度を計測
Perplexity(PPL)	TED(SI)により訓練された言語モデルで同時通訳文らしさを計測
BLEU	擬似同時通訳文における意味情報の保持具合を計測
RIBES	擬似同時通訳文の順送り文としての正確さを計測

表5. TEDコーパスの統計情報

Corpus	Train	Val	Test
TED(Caption)	21,239	1,000	1,000
TED(SI)	24,592	1,000	866

表6. 分類器精度及び言語モデルのPerplexity

二値分類器		言語モデル
Precision	**Recall**	**PPL**
98.09	98.09	59.83

[2] Tetsuji Nakagawa, Efficient Top-Down BTG Parsing for Machine Translation Preordering, 2015
[3] Guillaume Lample et al, Multiple-Attribute Text Rewriting, 2019

4)のコーパス対において同時通訳文らしい傾向

- 3)では単に2)と類似する文を生成する傾向
- 4)では2)の語順を保持しつつ流暢な同時通訳文を生成する傾向

表7. 自動評価指標による実験結果

	Corpus	ACC	PPL	BLEU	RIBES
1)	ASPEC(JA)	23.99	2733.59	55.71	62.12
2)	ASPEC(preordered)	32.47	3312.59	-	-
3)	ASPEC(JA) → ASPEC(preordered)	34.07	3360.39	**72.91**	81.88
4)	**ASPEC(preordered) →TED(SI) +CSJ + OS + JESC**	**72.32**	**782.43**	52.79	**91.01**

人手評価実験

自動評価指標のみを用いて擬似同時通訳文を評価するのは不十分

- 5人の被験者に30サンプルを提示

表8. 擬似同時通訳文の評価に用いた人手評価指標(1〜5点で評価)

Metrics	Detail
Word-order	英語原文と疑似同時通訳文の単語対応が一致している
Fluency	疑似同時通訳文は日本語として自然で流暢である
Colloquial	疑似同時通訳文は口語的である
Identity	日本語原文と疑似同時通訳文は意味的に同一である

4)のコーパス対において同時通訳文らしい傾向

- 3)は流暢さに欠いており非口語的
- Word-order, Fluency, Colloquialの評価指標において有意差あり(p<.001)

表9. 人手評価指標による実験結果

	Corpus Pair	Word-order	Fluency	Colloquial	Identity
3)	ASPEC(JA) → ASPEC(preordered)	3.46	2.90	2.24	**3.63**
4)	**ASPEC(preordered) → TED(SI) + CSJ + OS + JESC**	**3.64**	**3.47**	**3.37**	3.53

表10. 各評価指標における平均値が特に高かった擬似同時通訳文例
(Word-order=4.2, Fluency=4.4, Colloquial=4.6, Identity=4.6)

ASPEC(EN)	As one method for judging the pain mechanism, drug challenged test (DCT) is described.
ASPEC(JA)	痛みの機序を判定する1つの方法として薬理学的疼痛機序判別試験(DCT)について述べた。
ASPEC(preordered)	1つの方法として，痛みの機序を判定する薬理学的疼痛機序判別試験(DCT)について述べた
ASPEC(preordered) → TED(SI) + CSJ + OS + JESC	一つの方法として，痛みの機序を判定するための，薬理学的疼痛機序判別試験(DCT)について説明していきます。

表11. 各評価指標における平均値が特に低かった擬似同時通訳文例
(Word-order=2.6, Fluency=2.6, Colloquial=2.4, Identity=2.4)

ASPEC(EN)	Therefore, an approach to the problem from the art side by participation of artists was deceived, and the cooperative research was carried out by participation of an artist.
ASPEC(JA)	このため，アーティストが参加してアート側から問題にアプローチすることを考え，アーティストを加えて共同研究を行った。
ASPEC(preordered)	このため，考え，アーティストをことをアプローチする問題にアーティストが参加してアート側から共同研究を行った，加えて。
ASPEC(preordered) → TED(SI) + CSJ + OS + JESC	このため，アーティストを作ることにアプローチすることができますが，問題にアーティストが参加してアート側から共同研究を加えています 。