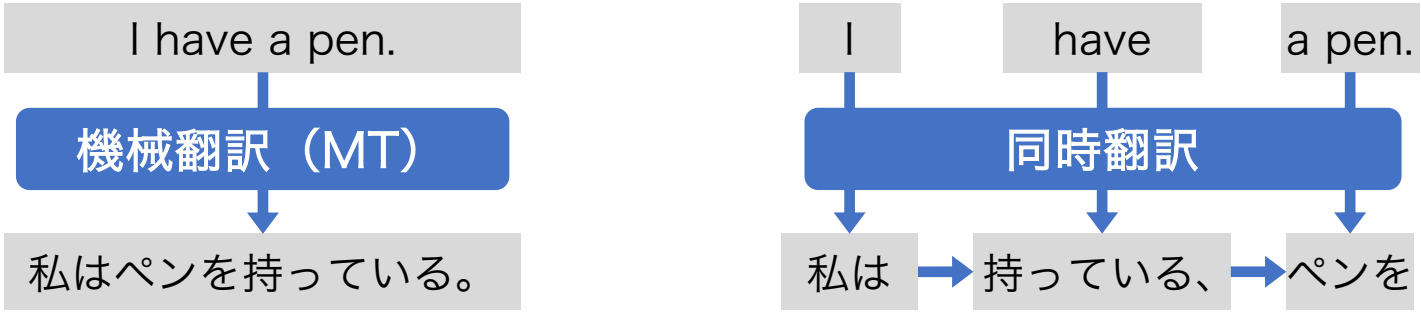
2024年3月10日 NL研

原発話に忠実な英日同時機械翻訳の実現に向けた順送り訳評価データ作成

福田 りょう, 土肥 康輔, 須藤 克仁, 中村 哲
奈良先端科学技術大学院大学

同時翻訳

- 文の終わりを待たずに翻訳を開始する漸進的な機械翻訳（MT）



- 同時翻訳の取り組み
  - 英語 → {ドイツ語、中国語、日本語} などの言語対で研究が盛ん
  - 多くは、時間に制約されない**オフライン翻訳**データを用いてMTを学習
    → 英日では語順の違いによって遅延が生じる（SVO→SOV）
  - **英日同時通訳**データを用いたMTの学習 [Shimizuら 2013, Koら 2023]
    → 通訳者の同時通訳の技法を取り入れることで低遅延な翻訳を目指す

同時通訳者の技法：順送りと短縮方略

英日通訳の同時性を高める2つの訳出方略 [遠山ら 2003]

1. 順送りによる訳出（**順送り方略**）
   - 発話をチャンクに区切りながら順次訳出
   - 時間的制約のため、順送り方略を常に適用することは難しい

> (1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.
>
> (1) アノニマスのようなグループが / (2) 台頭してきています，/ (3) 過去12ヶ月にわたって，/ (4) そして主要なプレイヤーになっています，/ (5) オンライン攻撃の分野において.

2. 短縮による訳出（**短縮方略**）
   - 要約や、重要でない箇所の省略などで遅延を低減

同時翻訳の課題：短縮方略の学習

- **同時通訳**データを用いたMTの学習 [Shimizuら 2013, Koら 2023]
  → 同時通訳の技法（順送り方略・短縮方略）の獲得を目指す

- MTが通訳者並みの適切さで短縮方略を行うことは難しい現状
  - 同時通訳事例を学習したMTは訳抜けが多い [Koら 2023]
  - 何を省略して何を省略しないかの区別が難しい. 多量のデータで解決？同時通訳データは少ない

原発話	• There’s no hospital that can say ”No.” • Anybody who’s paralyzed now has access to actually draw or communicate using only their eyes.
同時通訳	• 病院も、ノーとは言えない。 • 麻痺してる人達は、これを全員使うことが出来るようになっています。
同時通訳を学習したMT	• 病院は、 • 麻痺した人たちは、

目的

順送り方略だけを常に適用する同時翻訳システムの構築

- 順送り方略だけ … 省略しないことで、単語対応関係などが学習しやすい
- 常に適用 … 同時性が高く、かつ訳抜けの少ない（忠実な）同時翻訳を実現
  - 時間的制約から常に適用は難しかったが、処理が高速なMTなら可能？

	同時性	MTの学習	原発話への忠実さ
オフライン翻訳	×	○	○
同時通訳（順送りと短縮方略の使い分け）	○	×	△
順送り訳（順送り方略のみ）	○	○	○

本研究では初期検討として、MT評価用の英日順送り訳データを作成

関連研究：英日同時翻訳のデータ作成

1. 事前並び替えとスタイル変換による擬似同時通訳データ [二又ら 2020]
2. 文脈を考慮したチャンク単位の対訳コーパス [中林ら 2021]

↳ 自動的手法で大規模なデータを作成できる
訳文の流暢さに課題

3. 人手で作成した順送り訳を含む同時通訳コーパス [東山ら 2023]

↳ 本研究のデータと類似、かつ規模が大きい
本研究は (1) チャンキングを自動化, (2) 同時通訳と比較, (3) データを公開

順送り訳の定義

「チャンク単位で交差量が0、かつ省略が0となる訳出」

- 交差量：対応する訳文のチャンクの位置が右方向に移動した原発話のチャンクの数
- 省略：対応する訳文のチャンクが存在しない原発話のチャンクの数

原発話	(1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.	
オフライン翻訳	(1) Anonymousというグループは / (3) この12ヶ月ほど / (2) 活気付いていて / (5) オンライン攻撃において / (4) 大きな存在になってます	交差量=2 省略=0
同時通訳	(1) アノニマス集団ですね, こちらの方が, / (5) オンライン攻撃の, / (4) 主要な, プレイヤーになっています.	交差量=1 省略=2
順送り訳	(1) アノニマスのようなグループが / (2) 台頭してきています, / (3) 過去12ヶ月にわたって, / (4) そして主要なプレイヤーになっています, / (5) オンライン攻撃の分野において.	**交差量=0 省略=0**

順送り訳の定義

「チャンク単位で交差量が0、かつ省略が0となる訳出」

- 交差量：対応する訳文のチャンクの位置が右方向に移動した原発話のチャンクの数
- 省略：対応する訳文のチャンクが存在しない原発話のチャンクの数

原発話	(1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.	
オフライン翻訳	(1) Anonymousというグループは / (3) この12ヶ月ほど / (2) 活気付いていて / (5) オンライン攻撃において / (4) 大きな存在になってます	交差量=2 省略=0
同時通	マス集団ですね, こちらの方 **同時性↓** プレイヤーにな	交差量=1 省略=2
順送り訳	マスのようなグループが / (/ (3) 過去12ヶ月にわたって, / (4) そして主要なプレイヤー になっています, / (5) オンライン攻撃の分野において.	**交差量=0** **省略=0**

交差：(3)の完了を待って(2)を訳出

交差：(5)の完了を待って(4)を訳出

順送り訳の定義

「チャンク単位で交差量が0、かつ省略が0となる訳出」

- 交差量：対応する訳文のチャンクの位置が右方向に移動した原発話のチャンクの数
- 省略：対応する訳文のチャンクが存在しない原発話のチャンクの数

原発話	(1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.	
オフライン翻訳	(1) Anonymousというグループは / (3) この12ヶ月ほど / (2) 活気付いていて / (5) オンライン攻撃において / (4) 大きな存在になってます	交差量=2 省略=0
同時通訳	(1) アノニマス集団ですね, こちらの方が, / (5) オンライン攻撃の, / (4) 主要な, プレイヤーになっています.	交差量=1 省略=2
順送り訳	(1) アノニマスのようなグループが / (2) 台頭してきています, / (3) この12ヶ月に渡って, / (4) そして主要なプレイヤー / (5) オンライン攻撃の分野において.	**交差量=0** **省略=0**

(2)(3)が省略：原発話の情報全てが伝わらない

忠実さ↓

順送り訳の定義

「チャンク単位で交差量が0、かつ省略が0となる訳出」

- 交差量：対応する訳文のチャンクの位置が右方向に移動した原発話のチャンクの数
- 省略：対応する訳文のチャンクが存在しない原発話のチャンクの数

原発話	(1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.	
オフライン翻訳	(1) Anonymousというグループは / (3) この12ヶ月ほど / (2) 活気付いていて / (5) オンライン攻撃において / (4) 大きくなってます	交差量=2 省略=0
同時通訳	アノニマス集団ですね, こちらの方が, / (5) オンライン攻撃の, / (4) 主要な, プレイヤーになっています.	交差量=1 省略=2
順送り訳	(1) アノニマスのようなグループが / (2) 台頭してきています, / (3) 過去12ヶ月にわたって, / (4) そして主要なプレイヤーになっています, / (5) オンライン攻撃の分野において.	**交差量=0** **省略=0**

同時性と忠実さの観点で理想的

順送り訳の定義

「チャンク単位で交差量が0、かつ省略が0となる訳出」

- 交差量：対応する訳文のチャンクの位置が右方向に移動した原発話のチャンクの数
- 省略：対応する訳文のチャンクが存在しない原発話のチャンクの数

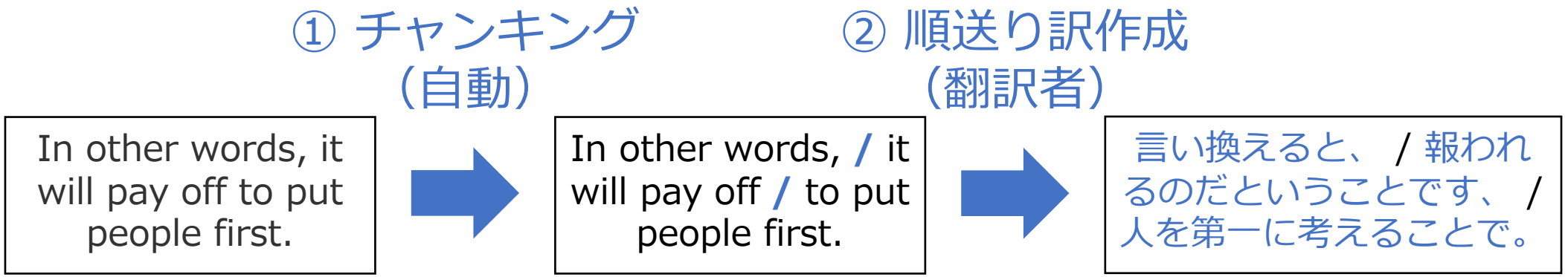
原発話	(1) Groups like Anonymous / (2) have risen up / (3) over the last 12 months / (4) and have become a major player / (5) in the field of online attacks.	
オフライン翻訳	(1) Anonymousというグループは(2)台頭してきていて / (3) 過去12ヶ月ほど / (4) 大きくなってます	交差量=2 省略=0
同時通訳	(1) アノニマスのような集団で(2)台頭してきていて / (3) 過去12ヶ月ほど / (4) オンライン攻撃の, / (4) 主要な, プレーヤーになっています.	交差量=1 省略=2
順送り訳	(1) アノニマスのようなグループが / (2) 台頭してきています, / (3) 過去12ヶ月にわたって, / (4) そして主要なプレイヤーになっています, / (5) オンライン攻撃の分野において.	**交差量=0** **省略=0**

同時性と忠実さの観点で理想的

つなぎ言葉による長い訳出
→ 聞き疲れしやすい？
一度の理解が難しい？

順送り訳データ作成：概要

- 手順



- 作業対象データ
  - TED talksに基づく英日同時通訳コーパスNAIST-SIC-Aligned-ST [Koら 2023] のテストデータ**511文**

順送り訳データ作成：① チャンキング

- 通訳者の経験に基づくチャンク化ルール [岡村, 山田　2023]
  - I.　節を導く接続詞，関係詞（主語を修飾する場合は除く）の前
  - II.　後ろに3語以上が続くto不定詞，前置詞，動名詞の前
  - III. 3語以上の長い主語の後
  - IV. コンマ（1語ずつの羅列は除く），セミコロン，ハイフン等の記号の前後
  - V.　文頭（または節を導く接続詞，関係詞の直後）の前置詞句，副詞句の後

I think / that
you are right.

- spaCyの英語モデルの構文解析情報を用いて自動化
  - 一貫性のあるチャンキング、翻訳者の作業負担軽減
  - 人手チャンキングとの一致：F1=75.9%, 適合率72.5%, 再現率79.7%

順送り訳データ作成：② 順送り訳作成

- 基本ルール
  1. 文頭からチャンクごとに翻訳する
  2. 文全体を見てチャンク間の繋がりが自然になるように翻訳する
  3. 後ろ（未来）のチャンクの情報を含めずに翻訳する

In other words, / it will pay off / to put people first.

△ 言い換えると, / それは報われるでしょう. / 人を第一に考えること.

○ 言い換えると, / 報われるのだということです, / 人を第一に考えることで.

つなぎ言葉

順送り訳データ作成：② 順送り訳作成

- 基本ルール
  1. 文頭からチャンクごとに翻訳する
  2. 文全体を見てチャンク間の繋がりが自然になるように翻訳する
  3. 後ろ（未来）のチャンクの情報を含めずに翻訳する

- 流暢性を保つために許容した操作（※下ほど慎重に適用）
  - 前のチャンクの情報を**繰り返し**訳出すること
  - 翻訳すべき情報を後ろのチャンクに**先送り**すること
  - 不要な情報を**省略**すること

順送り訳データ作成：② 順送り訳作成

- 基本ルール
  1. 文頭からチャンクごとに翻訳する
  2.
  3.

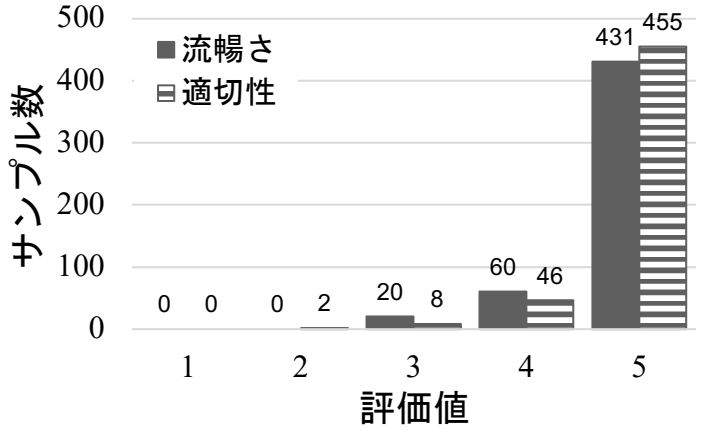
And he drew again / for the first time, / in front of his family / and friends.
△ そして彼は再び描きました, / 初めて, / 家族の前で / また友人たち.
○ そして彼は再び描きました, / 初めて, / 家族の前で / また友人たちの前で.

繰り返し

- 流暢性を保つために許容した操作（※下ほど慎重に適用）
  - 前のチャンクの情報を**繰り返し**訳出すること
  - 翻訳すべき情報を後ろのチャンクに**先送り**すること
  - 不要な情報を**省略**すること

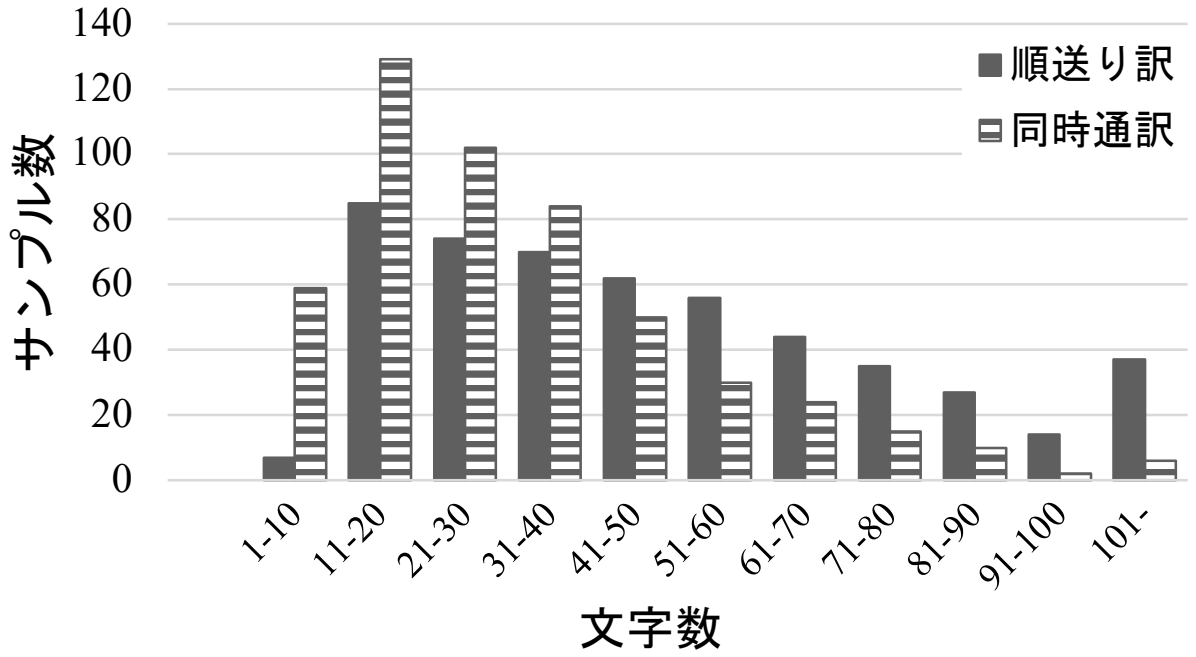
分析：品質評価

- 翻訳の評価基準で5段階評価（通訳経験のあるアノテータ）
  - 流暢さ：文章としてどの程度自然な表現か　→ 平均4.8/5
  - 適切性：原文の情報がどの程度含まれているか　→ 平均4.9/5
- 自然かつ原文の内容に忠実な順送り訳
  - 「全体にとても美しい順送り訳になっていると感じた」
  - システム構築の評価データとして適切



分析：同時通訳との比較

- オフライン翻訳を基準に、**同時通訳は短く、順送り訳は長い**
  - 合計文字数：同時通訳**1.6万**、オフライン翻訳**2.2万**、順送り訳**2.6万**
  - 文の平均文字数：同時通訳**31.6**、 順送り訳**50.3**



分析：同時通訳との比較

略により短く、順送り訳は繰り返しやつなぎる傾向

省略 “flexible design”
要約 “スクワットをしたりすること”

原発話	Its flexible design / allows for deep squats, / crawls and highagility movements.
同時通訳	これでスクワットをしたりすることができるわけです.
	計は / 深いスクワットを可能にし, / クロー す.

文脈を考慮した省略
“online attacks” → “攻撃”

繰り返し “allows

つなぎ言葉を多用
“のは, ” “です, それも”

原発話	They use online attacks / to make lots of money / and lots and lots of it
同時通訳	（直前の訳：オンラインの犯罪というのは，要するにお金を稼ごうとしてわけです．）これ攻撃をして，お金をたくさん稼いでると.
順送り訳	彼らがオンライン攻撃を用いるのは, / 多くのお金を稼ぐためです, / それも山ほどのお金を.

分析：不適切なチャンキング

- 自動チャンキングの精度やチャンキングのルールに改善の余地
  - 評価者「チャンク位置のために訳出困難，意味ずれなどが多発した」
  - 7%のサンプルに、チャンキングのエラーかエラー以外の不適切な分割

チャンキングのエラー
句動詞running intoの分割
→ 翻訳者は先送りで対処

ルールII*に従っているが
不自然な分割

*後ろに3語以上が続く前置詞

原発話	It was running / into bankruptcy last fall, / … … / along with most / of the other experimental quantum physicists, / …
順送り訳	それは， / 昨秋，破産寸前でした / … … / 大半の / 他の実験的量子物理学者たちと同じような類の一人であるという事です, / …

まとめと今後の課題

- **同時翻訳システム評価用の英日順送り訳データを作成**
  - 流暢さと適切性が高く、評価データとして適切な品質
  - チャンキングに改善の余地
  - データ：https://dsc-nlp.naist.jp/data/NAIST-SIC/Aligned-Chunk_Mono-EJ

- **聞き手の視点からデータを評価することが今後必要**
  - 長いほど一度に理解しづらい、聞き疲れしやすいなどの懸念を踏まえ「原発話に忠実な同時翻訳」の必要性を検証

- **その後、順送り訳を行える同時翻訳システムを構築**
  - 同時性が高く、原発話に忠実な同時翻訳の実現