When Does Translation Require Context? A Data-driven, Multilingual Exploration

Patrick Fernandes*, Kayo Yin*, Emmy Liu,
André F. T. Martins, Graham Neubig (ACL 2023)
(紹介者: NTT コミュニケーション科学基礎研究所 森下 睦)

概要

- 文脈を活用する機械翻訳モデルの研究が盛ん
- しかし、文脈翻訳モデルの評価はまだ発展途上
- そもそも翻訳時に文脈が必要なケースというのも

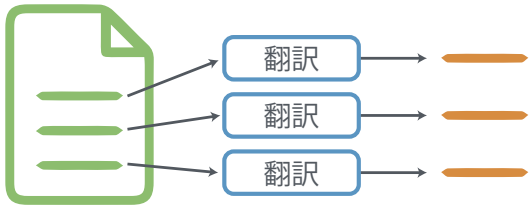
  言語によってはよくわかっていない
- 本論文のResearch Question
  - どのような場合に文脈翻訳を必要とする？
  - 現在の文脈翻訳モデルはどの程度それらを翻訳できる？
- ACL 2023 Best Resource Paper

背景

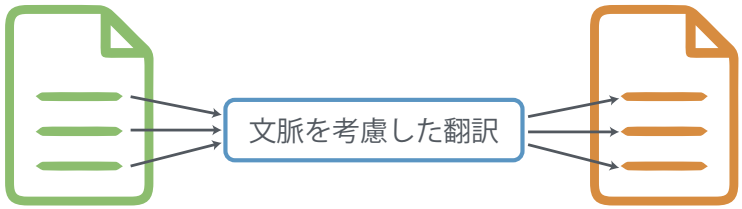
文脈を考慮した機械翻訳

- 通常機械翻訳では1文→1文への翻訳を行う

  それぞれの文は独立に翻訳



- 理想的には文脈（前後の文）を考慮して翻訳したい



例えば・・・

前文:　鈴木先生にお会いした。

入力文: すごい人だった。

翻訳文: He was a great person.

例えば・・・

前文:　今日は渋谷へ行った。

入力文: すごい人だった。

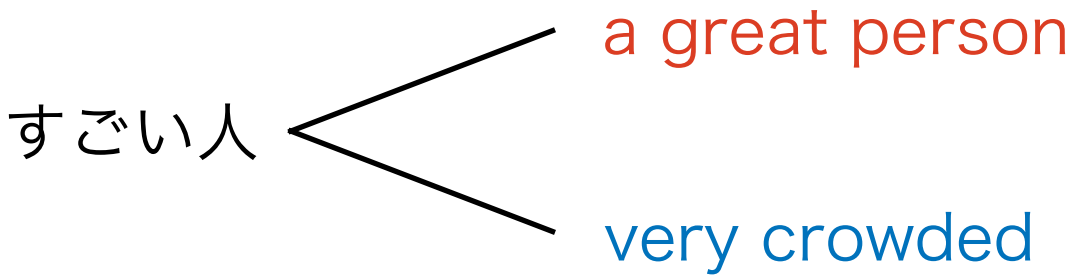
翻訳文: It was very crowded.

例えば・・・

前文:　今日は渋谷へ行った。

入力文: すごい人だった。

翻訳文: It was very crowded.



適切な翻訳文は文脈によって変化

→ 文脈を考慮できる翻訳器の研究が盛ん

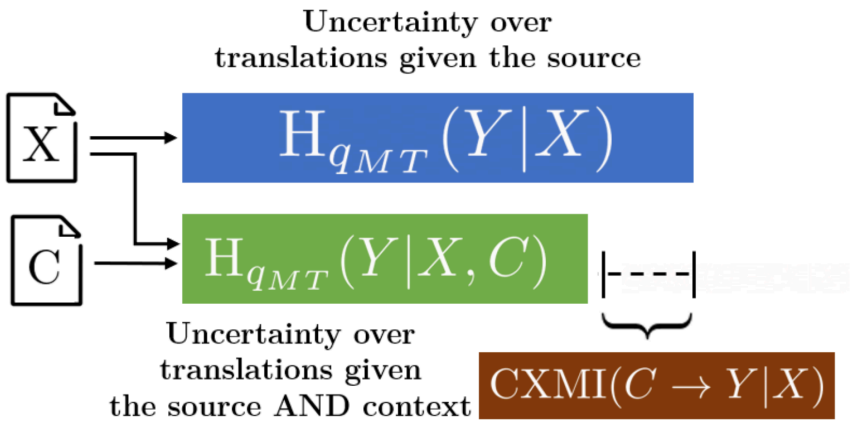
本論文の貢献

1. 文脈が必要な単語の自動検出法を提案
2. 文脈翻訳評価用データセットの作成
3. 既存の文脈翻訳モデルの再評価

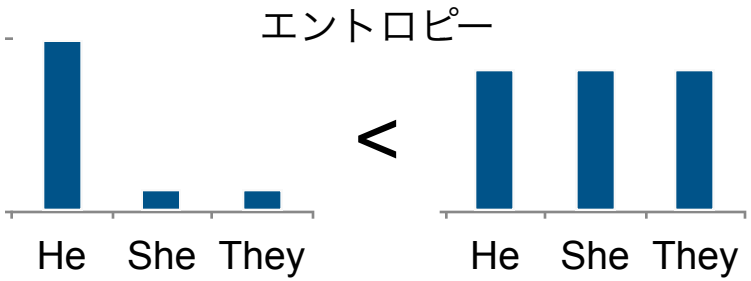
本論文の貢献

1. 文脈が必要な単語の自動検出法を提案
2. 文脈翻訳評価用データセットの作成
3. 既存の文脈翻訳モデルの再評価

先行研究:
Conditional Cross-Mutual Information (CXMI)



$$\mathrm{CXMI}(C \to Y|X)$$
$$= \mathrm{H}_{q_{MT_A}}(Y|X) - \mathrm{H}_{q_{MT_C}}(Y|X,C)$$
$$\approx -\frac{1}{N}\sum_{i=1}^{N} \log \frac{q_{MT_A}(y^{(i)}|x^{(i)})}{q_{MT_C}(y^{(i)}|x^{(i)}, C^{(i)})}$$



q_{MT_A}: 単文翻訳モデル

q_{MT_C}: 文脈翻訳モデル

- 機械翻訳モデルがどの程度文脈を活用しているかを計測
- CXMIが高い = モデルが文脈を活用できている

・Measuring and Increasing Context Usage in Context-Aware Machine Translation, Fernandes+, ACL 2021

・持橋さんのスライド P.9

提案法: Point-wise Conditional Cross-Mutual Information (P-CXMI)

文レベル

$$\text{P-CXMI}(y, x, C) = -\log \frac{q_{MT_A}(y|x)}{q_{MT_C}(y|x, C)}$$

単語レベル

$$\text{P-CXMI}(i, y, x, C) = -\log \frac{q_{MT_A}(y_i|y_{t<i}, x)}{q_{MT_C}(y_i|y_{t<i}, x, C)}$$

- 従来のCXMIを文・単語レベルに拡張
- P-CXMIが高い = 文脈を必要とする文・単語
- P-CXMIを活用することで、
  テストセット中の文脈を必要とする箇所を抽出可能に

いつ文脈が使われるのか

Avelile's mother had HIV virus. Avelile had the virus, she was born with the virus. 阿维利尔的母亲是携有艾滋病病毒。 阿维利尔也有艾滋病病毒。她一生下来就有。	Lexical Cohesion
Your daughter? Your niece? *Votre fille ?* Votre nièce ?	Formality (T-V)
Roger. I got'em. Two-Six, this is Two-Six , we're mobile. 了解 捕捉した。 2-6 こちら移動中だ。	Formality (Honorifics)
Our tools today don't look like shovels and picks. They look like the stuff we walk around with. *As ferramentas de hoje não se parecem com pás e picaretas.* Elas se parecem com as coisas que usamos.	Pronouns
Louis XIV had a lot of people working for him. They made his silly outfits, like this. *Luis XIV tenía un montón de gente trabajando para él.* Ellos hacían sus trajes tontos, como éste.	Verb Form
They're the ones who know what society is going to be like in another generation. I don't. *Ancak onlar başka bir nesilde toplumun nasıl olacağını biliyorlar.* Ben bilmiyorum.	Ellipsis

赤: P-CXMIが高い単語、 緑/青: 使われる文脈

- TEDトーク対訳コーパス (14言語) から
  P-CXMIが高い単語を分析し言語現象ごとに分類
- 主に5種類のパターンで文脈が使われることを発見
  - Verb Formなどはこれまで注目されてこなかったパターン

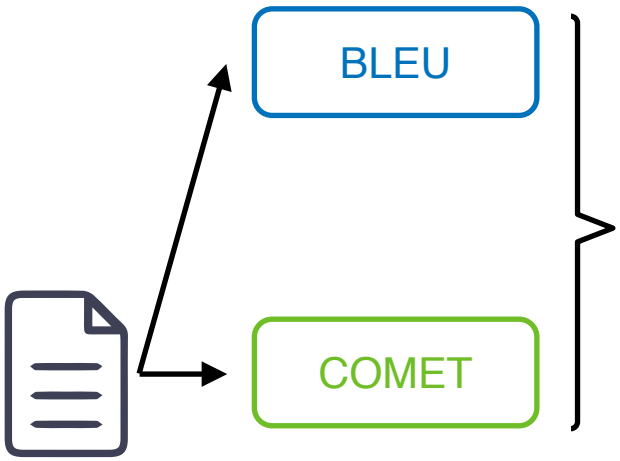
本論文の貢献

1 文脈が必要な単語の自動検出法を提案

2 文脈翻訳評価用データセットの作成

3 既存の文脈翻訳モデルの再評価

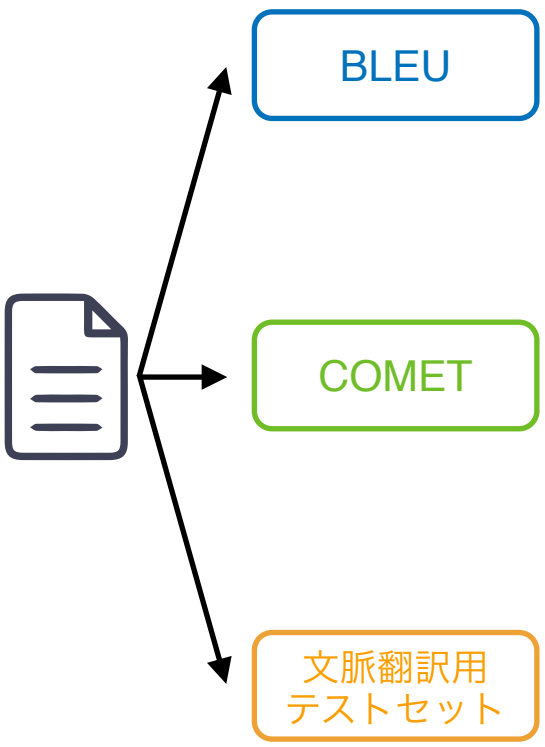
従来の文脈翻訳の評価



コーパスレベルの自動評価手法

- 従来から広く使われている評価手法
- コーパス全体のうち
  文脈を必要とする文・単語は少ない
- 出てきたスコアのみからは
  モデルが文脈を活用できているかは評価しにくい

従来の文脈翻訳の評価



コーパスレベルの自動評価手法

- 従来から広く使われている評価手法
- コーパス全体のうち
  文脈を必要とする文・単語は少ない
- 出てきたスコアのみからは
  モデルが文脈を活用できているかは評価しにくい

専用テストセットを使用する評価

- 文脈を必要とする文で構築されたテストセット
  - 各文の文脈を必要とする単語の翻訳精度を評価
- 一部の言語対・言語現象しか網羅していない
- テストセットの作成には言語知識が必要
  - これが自動で作れると理想

文脈が使われるケース

- 先程の分析によって
  以下のようなケースで文脈が使われることがわかった
  - Pronouns (代名詞)
  - Verb form (動詞の形)
  - Lexical cohesion (語彙の一貫性)
  - Formality (敬体)
  - Ellipsis (省略)
- コーパスからこれらの箇所を自動で抽出できれば、
  その翻訳精度を確認することで
  モデルが各現象にどの程度対応できているかわかる

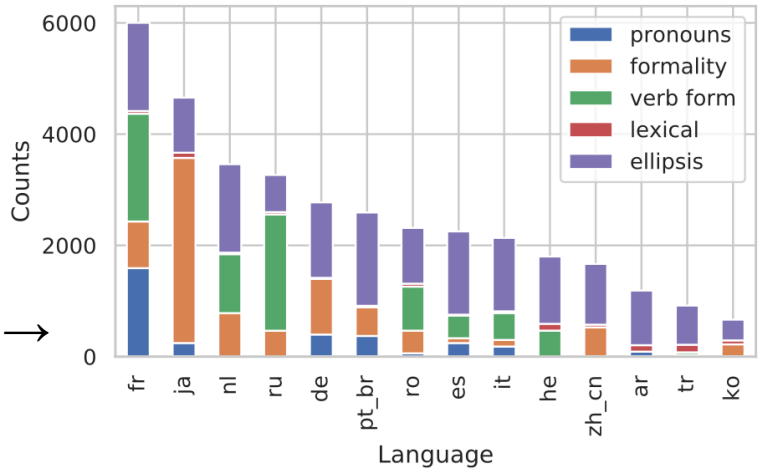
Multilingual Discourse-Aware (MuDA) benchmark

- 前述の言語現象を自動的に発見するタガーを構築
  - 詳細は割愛しますが、P-CXMIは使わずに

    言語現象ごとにルールや処理を考えて泥臭く構築
  - （これには言語特有の知識が必要なはず）
- TEDトークテストセット全体にタガーを適用し抽出

  → MuDAベンチマークと呼ぶ

日本語はFormalityが多い→



本論文の貢献

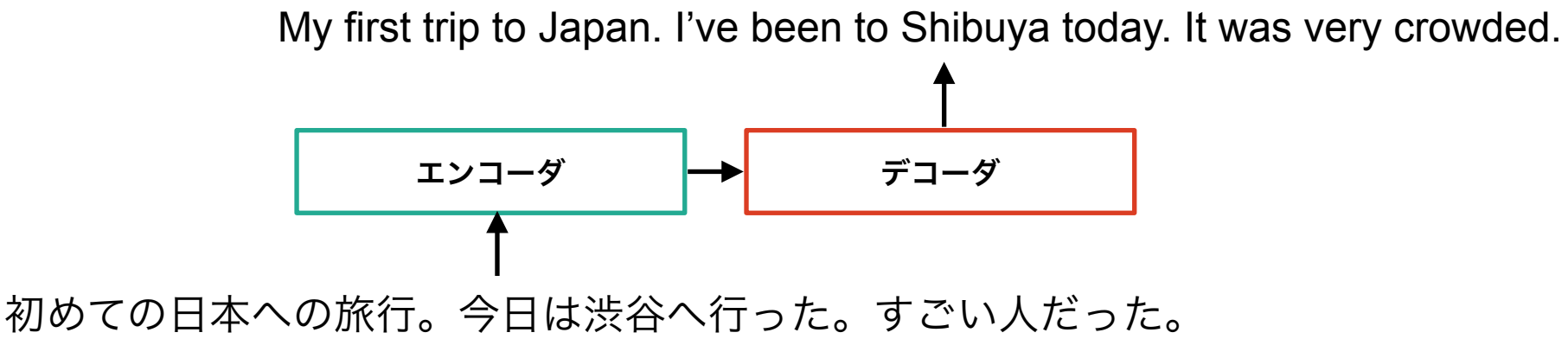
1 文脈が必要な単語の自動検出法を提案

2 文脈翻訳評価用データセットの作成

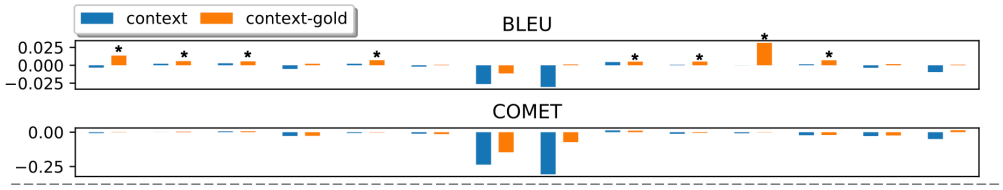
3 既存の文脈翻訳モデルの再評価

実験設定

- MuDAベンチマークを使って既存の文脈翻訳器を評価
  - どの言語現象ができていて、何ができていないかを調査
- 各言語について、単文翻訳モデル、文脈翻訳モデルを学習
- 文脈翻訳: 3-to-3モデル



実験結果



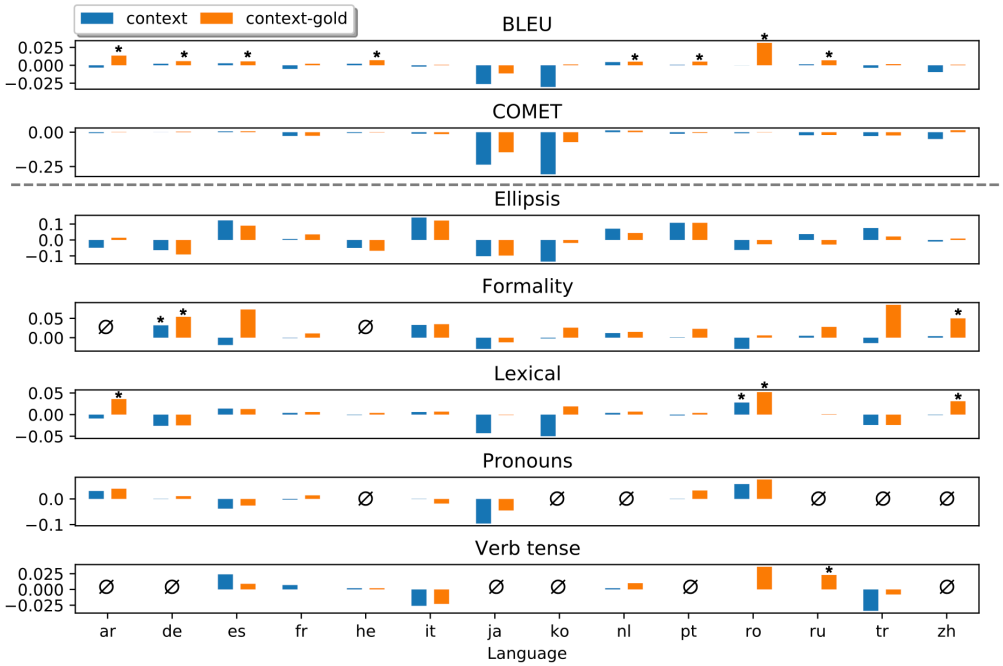
- BLEUとCOMETの傾向は不一致
  → 文脈が効いているのか
  よくわからない

単文翻訳モデルからの差分を表示

*: 単文翻訳モデルから統計的有意に精度向上

Context-gold: 文脈部分は正解文を入力

実験結果



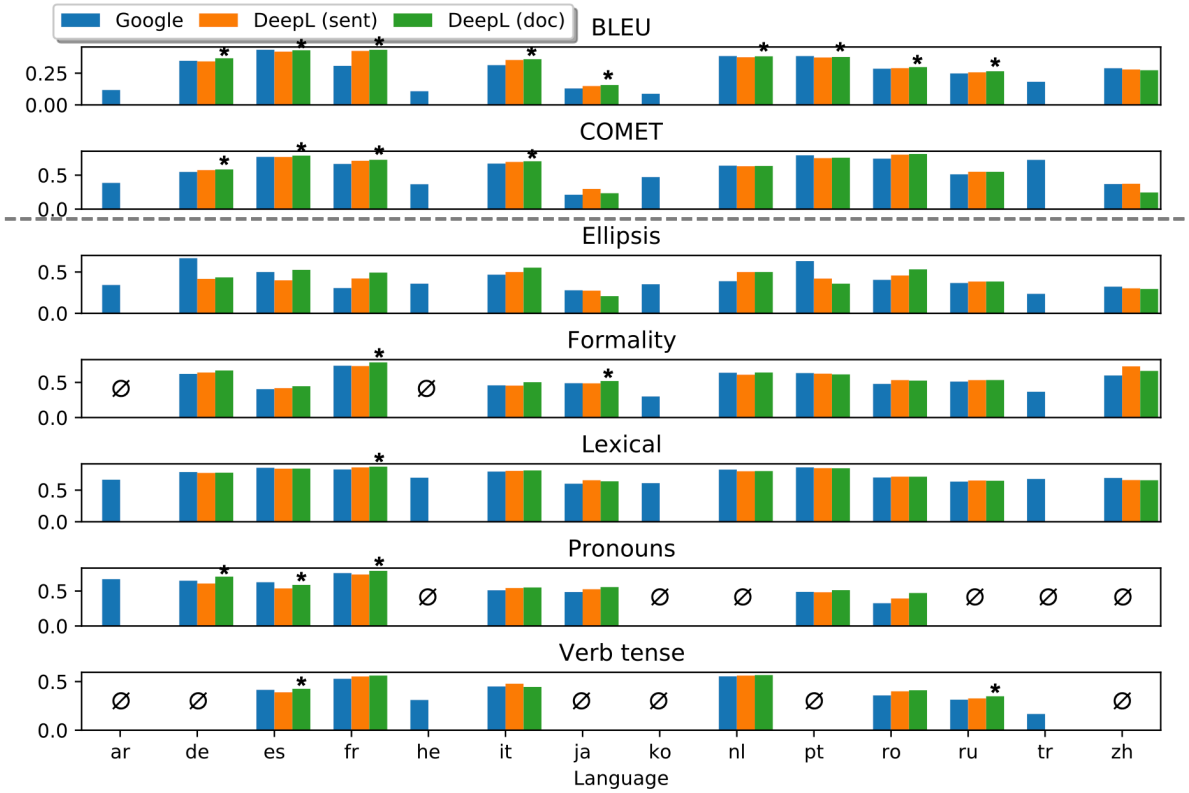
単文翻訳モデルからの差分を表示

*: 単文翻訳モデルから統計的有意に精度向上

Context-gold: 文脈部分は正解文を入力

- BLEUとCOMETの傾向は不一致
  → 文脈が効いているのか
  よくわからない
- 新たに作成したデータセット
  → 文脈翻訳モデルが一部の現象を正しく翻訳できていることがわかる
  ◯ Formality, Lexical cohesion
  × Ellipsis, Pronouns, Verb tense
- まだ翻訳できない言語現象も多く
  今後もモデルの改善が必要

商用システムの比較



- 文脈の活用では DeepL > Google ?
- DeepLは文脈を活用している (らしい) ので強い

まとめ

まとめ

1 文脈が必要な単語の自動検出法を提案

- P-CXMIを提案し14言語で文脈が必要なケースを分析
- 検出された単語を5種類のパターンに分類

2 文脈翻訳評価用データセットの作成

- 各言語現象を自動で検出するタガーを作成
- MuDAベンチマークを構築

3 既存の文脈翻訳モデルの再評価

- 現在の文脈翻訳モデルの課題を発見

所感

- 文脈を必要とする箇所をコーパス中から
  自動検出できるようになったのは良い話
- 一方で、詳細な分析を行うためには結局言語特有の知識が必要で、評価を完全に自動化できるかというと難しい
- 最近のLLMを使った翻訳は、長い文脈を使った文の翻訳に強いらしいので、それも比較対象に加えてもらえると今何が課題なのかわかって嬉しそう

参考文献

- 著者らのポスター、スライド
  - https://virtual2023.aclweb.org/paper_P3784.html

END