A7-4

2024.3.13
言語処理学会第30回年次大会

同一の原文書に対する複数の翻訳文書間で対応する言語単位対の自動抽出

本田友乃（東京大学）

藤田篤（NICT）

本研究の概要

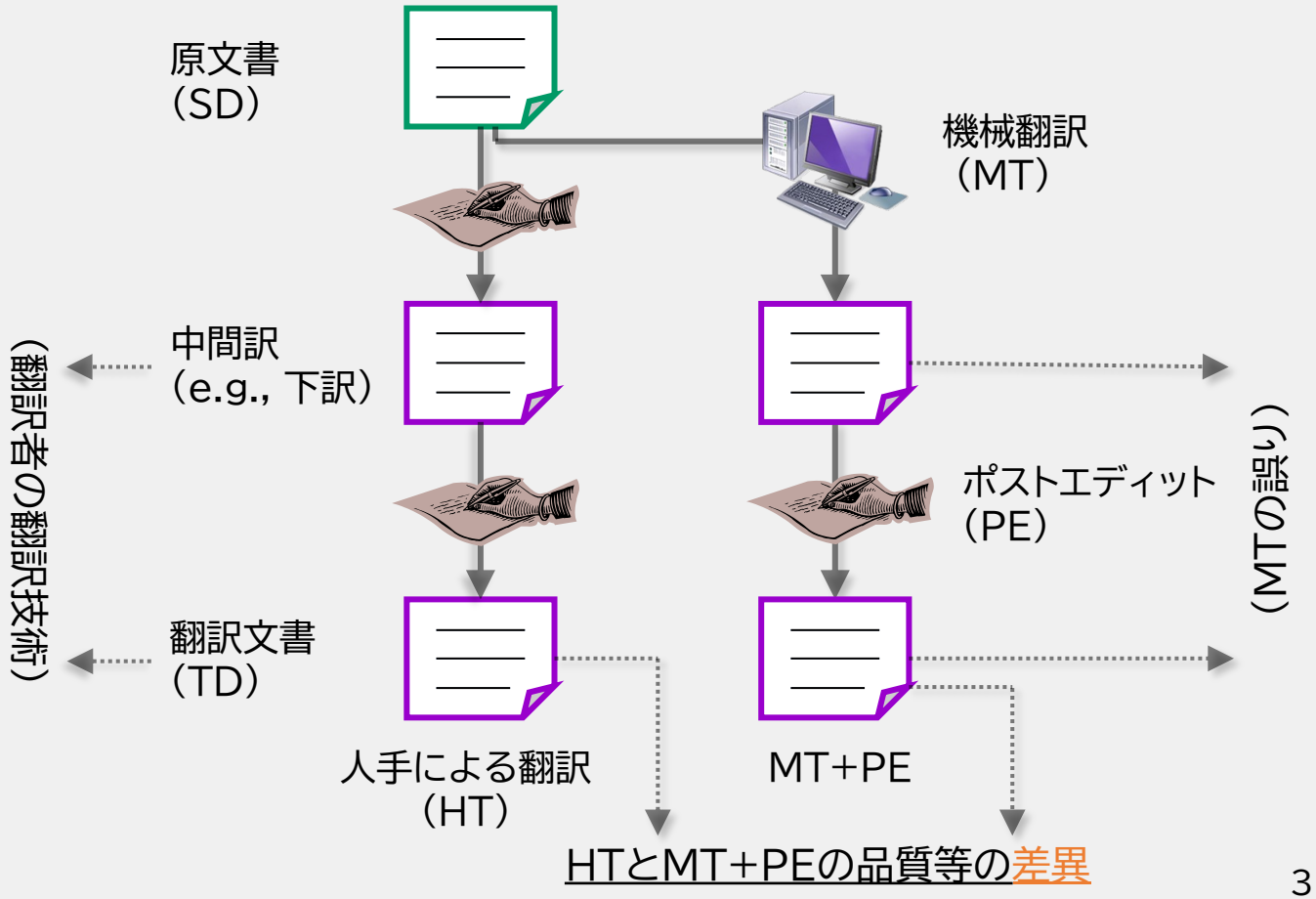
- 同一の原文書に対する複数の翻訳文書間で対応する言語単位対の自動抽出
  - 分割元の言語単位の組に内包される階層的な言語単位対を網羅的に抽出
  - 誤りのない翻訳間の分析に向けた効率化が目的
  - 単語対からフレーズ対をボトムアップに生成する手法を提案

原文書の言語単位	翻訳文書Aの言語単位	翻訳文書Bの言語単位
You are required to pay the required fee.	必要な手数料を納付しなければならない。	所要の料金の支払いが求められます。
the required fee	必要な手数料	所要の料金
required	必要な	所要の
fee	手数料	料金
You are required to pay	納付しなければならない	支払いが求められます
.	。	。

・入れ子のない単位に分割
・妥当な言語単位のみを抽出

翻訳間の差異の分析

- 異なる工程で産出された誤りのない翻訳間に見られる差異の分析
  - e.g., 異なる翻訳者による翻訳、
    人手による翻訳(HT)と
    ポストエディット(MT+PE)
  - 各工程の質的な違いを
    理解する上で重要



差異を分析するためのスキーム

- 英日翻訳を対象とした人手分析用のスキーム(本田ら, 2022; Honda et al., 2023)
  - 翻訳文書対から言語単位対への**分割**
    - 入れ子のない要素への分割: 分割元の単位の統語的な構造に従って、
      (A)分割後の要素同士に重なりがないように分割する
      (B)できるだけ大きな単位を抽出する
      という2つの条件を満たす操作
  - 各言語単位対における差異の**分類**
- このスキームを用いた人手分析には、効率化が必要
  - e.g., 450ワード程度の原文書に対する翻訳文書対の分割には6時間程度必要

タスクの定義

- 入力: <SD, TD_1, TD_2>の組
- 出力: 入力の組に内包されている階層的な言語単位の組の集合
  $\{(s, t_1, t_2) \mid s \sqsubset SD, t_1 \sqsubset TD_1, t_2 \sqsubset TD_2, s \equiv t_1 \equiv t_2\}$
- 補足
  - SD : 翻訳の原文書(Source Document)
  - TD_1、TD_2 : 当該SDに対する異なる翻訳文書(Target Document)
  - $x \sqsubset X$: x が文書Xに内包される要素である
  - $\equiv$: 当該文書において文書要素が対応している
  - SD及び$s \sqsubset SD$は任意

使用データ

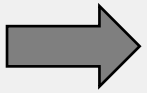
- MultiEnJaを利用
  - 誤りのない翻訳間の分析用に作られたデータセット
- 開発用データと評価用データを1種類ずつ準備
  - 開発用データ: 提案手法の検討
  - 評価用データ: 提案手法の性能評価
- 人手分析用のスキームの評価実験の結果をもとに正解データを作成

データの種別	単位対数	文数	語数(トークン数)		
			SD	HT	MT+PE
開発用データ	328	37	378	520	515
評価用データ	434	47	421	591	570

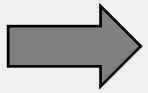
提案手法の概要

- 多様な対を得ることを目的として、語の対から大きい単位対をボトムアップに組み合わせるアプローチを検討

TD1	TD2
本	この
ページ	ページ
の	の
対象	対象
者	対象
:	:



TD1	TD2
本	この
本ページ	このページ
本ページの	このページの
本ページの対象者	このページの対象
本ページの対象者:	このページの対象:
ページ	ページ
⋮	

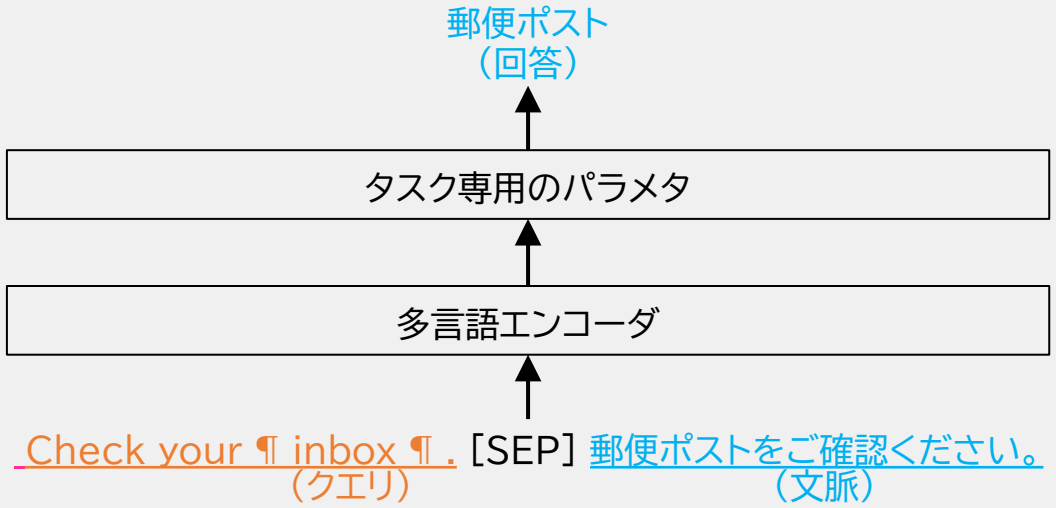


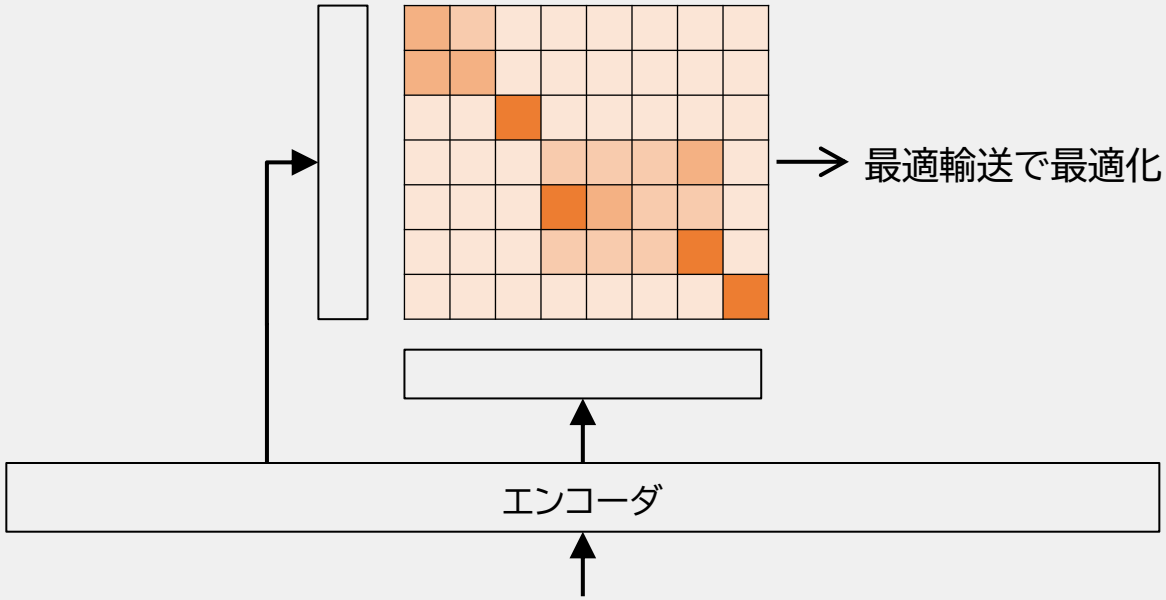
TD1	TD2
本ページの対象者	このページの対象
本ページ	このページ
本	この
ページ	ページ
の	の
対象者	対象
:	:

1. 単語対の取得
2. フレーズ対の取得
3. フレーズ対の組み合わせ
   （ステップ2の出力の和集合の取得）
4. 後処理(フィルタリング)

1. 単語対の取得

- アプローチの異なる単語アラインメントツールを利用して4種類の単語対を取得
- WSPAlign (Wu et al., 2023)
  - SQuAD形式の質問応答タスク(Rajpurkar et al., 2016)を応用
  - それぞれの単語に対して、クエリに対する回答(文脈中のスパン)を独立に求める
- OTAlign (Arase et al., 2023)
  - 入力対の単語埋め込みから全体の対応づけを求める





1. 単語対の取得

- 4種類のTD間の単語対を取得
  a. SDを介したTD間の単語対（WSPAlignを使用）
     - <SD, TD_1>と<SD, TD_2>に対してそれぞれWSPAlignを適用し、SD-TD間の単語対を得た上で、SD中の同じ単語に対応づけられたTD_1、TD_2の単語を対応づけることでTD間の単語対を取得
  b. TD間の単語対（WSPAlignを使用）
  c. TD間の単語対（OTAlignのUOTを使用）
  d. TD間の単語対（OTAlignのPOTを使用）

2. フレーズ対の取得

- ステップ1（単語対の取得）で得たa～dそれぞれの単語対の集合から
  フレーズ対を取得
  - 統計的機械翻訳におけるフレーズ抽出(Och et al., 1999)のアルゴリズムを適用
  - Moses Phrase extractor (Koehn et al., 2007)を利用

3. フレーズ対の組み合わせ

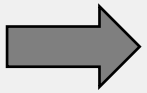
- ステップ2（フレーズ対の取得）で得た結果を組み合わせる
  - a～dの4通り → 15通り
- フレーズ対を組み合わせた結果に対して、それぞれTD間で表層上一致する単語対を追加
  - ステップ1～2では表層文字列が考慮されていないため、TD間で表層上一致する単位対が漏れている可能性が高い
  - Meteor（Denkowski & Alon Lavie, 2014）を利用
    - ひらがな一文字同士の単語対を除く

4. 後処理

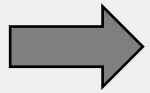
- 句構造解析器を用いた絞り込み
  - ステップ3（フレーズ対の組み合わせ）の結果から、下線部のような文法的に妥当ではない単位を除外する
    - e.g., 現在、更新通知は電子メールで送信されています。
  - Jigg（Noji & Miyao, 2016）を利用
- 分割元の単位対が表層上一致する対の除外
  - 人手分析用のスキームにおける分割手順で定められた分割停止条件に対応

提案手法の概要

TD1	TD2
本	この
ページ	ページ
の	の
対象	対象
者	対象
:	:



TD1	TD2
本	この
本ページ	このページ
本ページの	このページの
本ページの対象者	このページの対象
本ページの対象者:	このページの対象:
ページ	ページ
⋮	



TD1	TD2
本ページの対象者	このページの対象
本ページ	このページ
本	この
ページ	ページ
の	の
対象者	対象
:	:

1. 単語対の取得

2. フレーズ対の取得

3. フレーズ対の組み合わせ
   （ステップ2の出力の和集合の取得）
4. 後処理(フィルタリング)

評価実験

- ステップ3（フレーズ対の組み合わせ）で得られる15通りの全ての結果に対して正解データに対する再現率、適合率、F1値を算出

色付きの数値は予稿から修正したもの

評価結果(修正版)

評価対象	開発用データ			評価用データ		
	再現率	適合率	F1値	再現率	適合率	F1値
a. SDを介したTD間の単語対(WSPAlign)	52.1	68.7	59.3	33.9	55.1	41.9
b. TD間の単語対(WSPAlign)	55.5	**74.9**	63.8	53.0	**69.5**	**60.1**
c. TD間の単語対(OTAlign、UOT)	42.7	61.7	50.5	39.6	53.1	45.4
d. TD間の単語対(OTAlign、POT)	42.1	64.2	50.9	34.3	54.6	42.1
a･b	65.2	69.0	**67.1**	56.5	59.5	57.9
a･c	61.0	63.5	62.2	50.7	49.8	50.2
a･d	59.8	64.7	62.1	49.1	51.0	50.0
b･c	60.4	68.5	64.2	57.1	58.9	58.0
b･d	60.1	70.4	64.8	56.0	60.6	58.2
c･d	43.6	60.9	50.8	40.8	51.5	45.5
a･b･c	**67.4**	64.2	65.8	59.7	52.5	55.9
a･b･d	67.1	65.5	66.3	59.2	54.1	56.6
a･c･d	61.0	62.7	61.8	52.1	48.8	50.4
b･c･d	60.4	67.6	63.8	57.6	56.9	57.2
a･b･c･d	**67.4**	63.5	65.4	**60.1**	51.2	55.3

色付きの数値は予稿から修正したもの

評価結果(修正版)

評価対象	開発用データ			評価用データ		
	再現率	適合率	F1値	再現率	適合率	F1値
a. SDを介したTD間の単語対(WSPAlign)	52.1	68.7	59.3	33.9	55.1	41.9
b. TD間の単語対(WSPAlign)	55.5	**74.9**	63.8	53.0	**69.5**	**60.1**
c. TD間の単語対(OTAlign、UOT)	42.7	61.7	50.5	39.6	53.1	45.4
d. TD間の単語対(OTAlign、POT)	42.1	64.2	50.9	34.3	54.6	42.1
a･b	65.2	69.0	**67.1**	56.5	59.5	57.9
a･c	61.0	63.5	62.2	50.7	49.8	50.2
a･d	59.8	64.7	62.1	49.1	51.0	50.0
b･c	60.4	68.5	64.2	57.1	58.9	58.0
b･d	60.1	70.4	64.8	56.0	60.6	58.2
c･d	43.6	60.9	50.8	40.8	51.5	45.5
a･b･c	**67.4**	64.2	65.8	59.7	52.5	55.9
a･b･d	67.1	65.5	66.3	59.2	54.1	56.6
a･c･d	61.0	62.7	61.8	52.1	48.8	50.4
b･c･d	60.4	67.6	63.8	57.6	56.9	57.2
a･b･c･d	**67.4**	63.5	65.4	**60.1**	51.2	55.3

a･b･c･dの組み合わせについて
ステップ4の後処理の前後の性能を比較
- 再現率: 81.6%から60.1%に低下
- 適合率: 3.5%から51.2%に改善

エラー分析

- 再現率が最も高かった、a・b・c・dすべてのフレーズ対を組み合わせた場合の抽出結果を対象として、抽出もれ・抽出誤りを分析

色付きの数値は予稿から修正したもの

抽出もれ

		MT+PE							
		文相当	節	語句相当	句	複合表現	語	記号	対応なし
HT	文相当	-	-	-	-	-	-	-	-
	節	-	5/17	-	0/1	-	-	-	-
	語句相当	-	-	3/11	-	-	-	-	-
	句	-	0/2	-	12/40	3/8	1/1	-	-
	複合表現	-	-	-	9/10	38/78	11/17	1/1	0/3
	語	-	-	-	2/3	4/6	128/157	0/1	0/5
	記号	-	-	-	-	-	-	44/52	0/5
	対応なし	-	-	-	-	0/3	0/5	0/8	-

色付きの数値は予稿から修正したもの

抽出もれ

		MT+PE							
		文相当	節	語句相当	句	複合表現	語	記号	対応なし
HT	文相当	-	-	-	-	-	-	-	-
	節	-	5/17	-	0/1	-	-		
	語句相当	-	-	3/11	-	-	-		
	句	-	0/2	-	12/40	3/8	1/1	-	-
	複合表現	-	-	-	9/10	38/78	11/17	1/1	0/3
	語	-			2/3	4/6	128/157	0/1	0/5
	記号	-			-	-	-	44/52	0/5
	対応なし	-	-	-	-	0/3	0/5	0/8	-

一方が「対応なし」となる単位対

比較的大きな単位対

抽出もれ

- 対の一方が「対応なし」となる単位対
  - e.g., <**承認可能な**><支払い方法>　　<支払方法>
  - 語の対からフレーズ対を組み合わせるという方針では原理的に抽出できない
- 比較的大きな単位対（節、語句相当、句、複合表現）
  - e.g., <**建物サービスの各クラスに対する所定の要件**><は><、><更新用の用紙><に><記載されます><。>
<更新申請書><では><、><**各建築工事区分に対する規定の要件**><が>
<定められています><。>
  - ステップ1（単語対の取得）とステップ2（フレーズ対の取得）だけでは同定が困難
    - ステップ4（後処理）で除外してしまっているわけではない
  - フレーズ対を直接同定する手法を検討中
    - e.g., フレーズアラインメント（Arase & Tsujii, 2020）、WSPAlignのファインチューニング

色付きの数値は予稿から修正したもの

抽出誤り

i. 単位・対ともに適切だが不要な対: 91件

ii. 単位は適切だが対が不適切: 134件

iii. 一方または双方の単位が不適切: 24件

- i・ii: 句の単位対
  - 「入れ子のない要素への分割」と句構造解析器による解析が一致しない場合が多い
  - e.g., <あなたの投稿ボックス><ではなく><受信トレイ><を><チェックしてください><:>
<郵便ポスト><ではなく><受信トレイ><を><ご確認ください><。>
    - i:「受信トレイをチェックしてください」と「受信トレイをご確認ください」
    - ii:「チェックしてください」と「受信トレイをご確認ください」
- iii: 記号を含む単位対
  - 記号を先頭に置く単位など、文法的に妥当ではない単位
  - e.g., <個人><-><用紙52>　　<個人><ー><書式52>

まとめ

- 同一の原文書に対する複数の翻訳文書間で対応する言語単位対の自動抽出
  - 誤りのない翻訳間の差異の分析に向けた効率化が目的
  - 単語対からフレーズ対をボトムアップに生成する手法を提案
- 提案手法の性能
  - 再現率: 56.5%、適合率: 59.5%、F1値: 57.9% (a・bの組み合わせ)
  - 比較的大きな単位対の抽出が特に困難
- 今後の展望
  - フレーズ対の直接の取得
    - e.g., WSPAlignのファインチューニング、フレーズアラインメントツールの利用