NLP J Anthology BTPPコーパス:自然言語処理分野の論文調査支援を目的とした英日翻訳後処理コーパスの作成

[S1-P03] 中町礼文 西原大貴 （janthology.jp）
{aka_underfirst, skatenosenshu}@janthology.jp

論文調査支援ツール
https://janthology.jp

目的：NLP初学者の論文調査支援に向けて英語論文を和訳したい

- NLPの非専門家にとって、年々増加するACL論文を英語で読むのは大変
- 既存の汎用機械翻訳(MT)モデルは専門用語やスタイルを正しく訳せない →誤り訂正したい

入力文	汎用MT出力文	正解文
… finetuned …	… 微調整しました。	… ファインチューニングした。
Text simplification is …	テキスト単純化は …	テキスト平易化は …

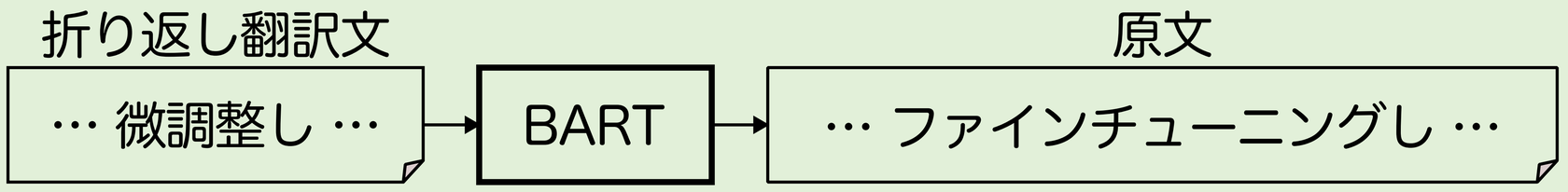
手法：折り返し翻訳で訓練データ作成（公開予定・貢献①）し、BARTで後編集

Step 1: 日本語論文から日→英→日の折り返し翻訳文コーパスを作成

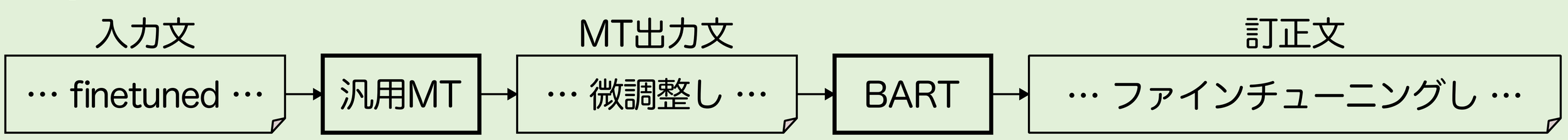
作成したコーパスの例（FuguMT [1] で折り返し翻訳）

折り返し翻訳文	原文
実際、日本の側が統計機械翻訳(SMT)によって翻訳される翻訳サイトです(第3.2)。	実際には, 日本語側が統計的機械翻訳 (SMT) によって翻訳された対訳サイトである (3.2 節参照).

Step 2: BARTを訓練



Step 3: 訓練したBARTで翻訳文を誤り訂正



実験：作成したコーパスにより翻訳性能改善（貢献②）

汎用MT：FuguMT [1]

後編集：京大BART [2] (baseline)
AcademicBART [3] (学術論文BART)

データ：年次大会21〜24年の概要 [4] 8,034件
訓練:検証:評価 = 8:1:1 に分割

前処理：bunkai [5] で文分割し、
文字数と編集距離でフィルタ

自動評価（提案手法でBLEUが8.00改善）

	BLEU	TER
汎用MT	43.72	37.25
汎用MT + 京大BART$_{base}$	48.86	34.74
汎用MT + 京大BART$_{large}$	49.51	34.91
汎用MT + AcademicBART	**51.72**	**31.86**

出力例（専門用語やスタイルが改善）

入力文	汎用MT出力文	BART出力文
It is worthwhile to assist dentists with such technology to prevent errors by inexperienced dentists and to reduce the workload of experienced ones.	経験の浅い歯科医によるエラーを防止し、経験豊富な歯科医の負担を軽減するために、そのような技術で歯科医を支援する価値があります。	経験の浅い歯科医によるエラーを防止し, 経験豊富な歯科医の負担を軽減するために, このような技術の活用は, 歯科医支援の有効な手段となる.
Controllable text simplification assists language learners by automatically rewriting complex sentences into simpler forms of a target level.	制御可能なテキスト単純化は、複雑な文をターゲットレベルの単純な形式に自動的に書き換えることで、言語学習を支援する。	制御可能なテキスト平易化は, 複雑な文をターゲットレベルの単純な形式に自動的に書き換えることで言語学習を支援する.

[1] https://github.com/s-taka/fugumt [2] https://nlp.ist.i.kyoto-u.ac.jp/?BART日本語Pretrainedモデル [3] 山内ら (2023) 学術ドメインに特化した日本語事前訓練モデルの構築, 第29回年次大会 [4] https://huggingface.co/datasets/kunishou/J-ResearchCorpus [5] https://github.com/megagonlabs/bunkai