JParaBank: 機械翻訳に基づく大規模な言い換え文対の収集

樽本空宙、惟高日向、梶原智之、二宮崇（愛媛大学）

1. 概要

- 日本語の言い換え文生成に基づくデータ拡張
- 言い換えモデルを訓練するためのデータセットは、英語や中国語では存在するが日本語には無い
- 言い換え文対を約2,100万文対収集したデータセットを構築、公開　　https://github.com/EhimeNLP/JParaBank

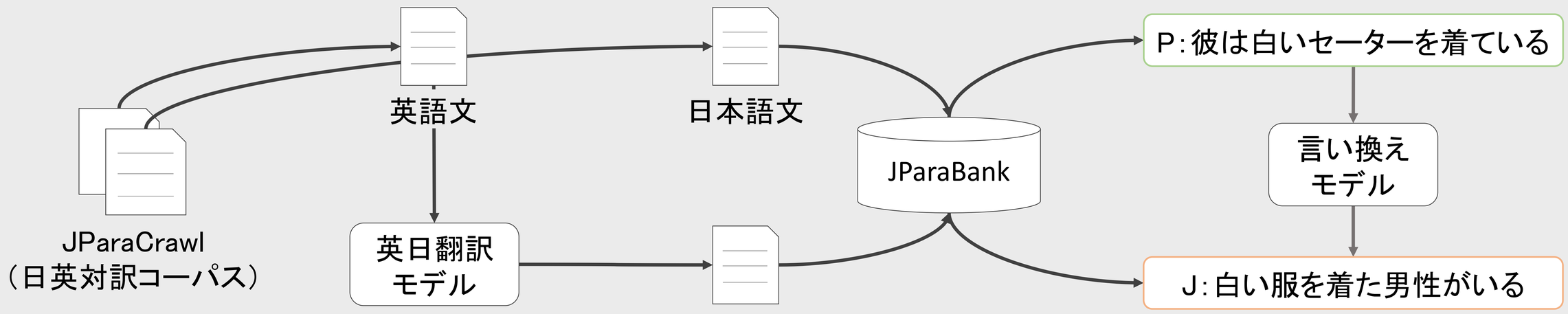
2.言い換え文対の例

原文	言い換え文
多くの大型望遠鏡や天文台は、高い山の上にあります。	多くの大きな望遠鏡や天文台は高い山頂にあります。
入力した場所以外の部分をクリックして確定します。	入力以外の場所をクリックして確定します。
コクのあるリッチな使い心地のクリームタイプ。	使用感が豊かなクリームタイプ。

単語の言い換えや追加、削除だけでなく、文全体を言い換えた文対が含まれる

3.JParaBankの構築と言い換えモデルの訓練

1. 大規模な日英対訳コーパスJParaCrawlのうち、英語文を日本語文に翻訳
2. JParaCrawlの日本語文（J）と翻訳した日本語文（P）を言い換え文対として収集、一致するJとPを削除しJParaBankを構築
3. JParaBankのうち、PからJの方向で言い換えモデルを訓練



日本語の言い換え文

4. 内的評価：言い換えモデルが生成した文の品質を人手評価

- 日本語言語理解ベンチマークJGLUEの各タスクから無作為に20文対ずつ抽出した100文対に対して評価
- 意味の等価性と文の流暢性を5段階評価
- 日本語母語話者の大学生3名による評価

	意味の等価性	文の流暢性
折り返し翻訳	2.67	3.25
JParaBank（J→P）	4.18	4.49
JParaBank（P→J）	4.62	4.71

提案手法によって生成された言い換え文の品質が最も高かった

5.外的評価：日本語言語理解ベンチマークJGLUEによる評価

- JGLUE：感情極性分類、文類似度推定、含意関係認識、質問応答で構成されるベンチマーク
- いずれのデータ拡張手法もデータ量を2倍に拡張

東北大BERT_base	MARC-ja (acc)	JSTS (pearson)	JNLI (acc)	JSQuAD (F1)	JCQA (acc)
データ拡張なし	0.953	0.903	0.887	0.935	0.805
ノイズ付与(EDA)	0.952	**0.905**	**0.895**	**0.936**	0.796
折り返し翻訳	**0.954**	**0.906**	0.866	0.933	**0.810**
JParaBank	**0.956**	**0.906**	**0.896**	**0.937**	**0.809**

提案手法によるデータ拡張が最も多くのタスクでベースラインの性能を上回った