P8-16

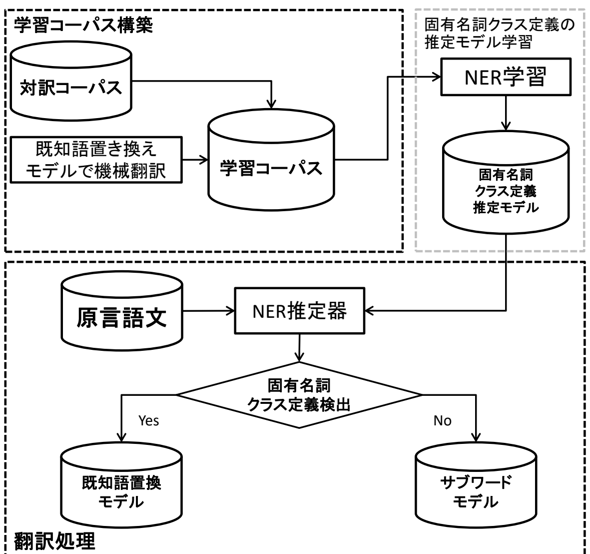
翻訳精度に基づく固有名詞の翻訳手法の研究

高井 公一†‡、服部 元‡、米山暁夫‡、安田 圭志†、須藤克仁†、中村哲†
†奈良先端科学技術大学院大学　‡株式会社 KDDI総合研究所

背景

〇 ランドマーク、飲食店、宿泊施設などの固有名詞数多く存在し、そのカバレージが**翻訳システムの精度に影響**
〇 固有名詞の正確な翻訳の解決手法の一つに、クラス言語モデル手法がある
〇 固有名詞の対訳辞書の作成コストと固有名詞クラスの**アノテーションのコストが問題**
〇 クラス言語モデルに必要な**アノテーションの課題を解決**し、**固有名詞を正確に訳せる機械翻訳**を実現

全体処理

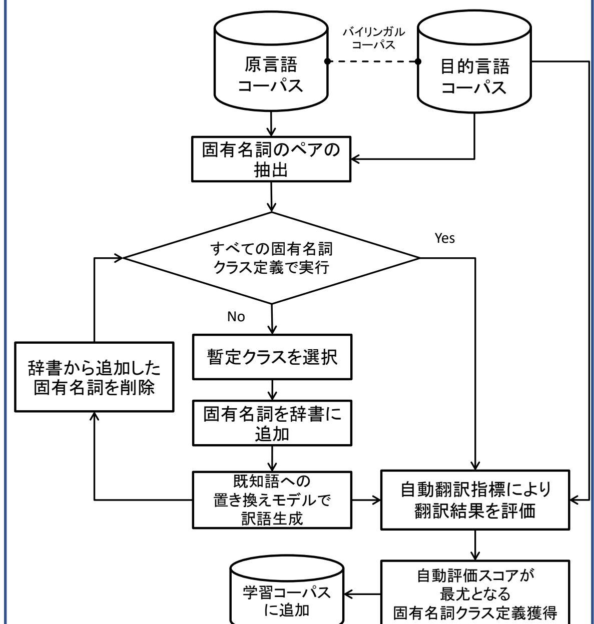


固有名詞を正確に翻訳する３パート

(1) アノテーションなしの学習コーパス構築
(2) 従来手法３つで複数学習
(3) 原文の固有名詞を検出、ハイブリッド翻訳処理

※ハイブリッド翻訳
■NERで固有名詞のある文は、推定した固有名詞クラスの代表単語に置換して、翻訳処理

学習コーパス構築



アノテーション作業のないNER学習データの構築手順

1. 固有名詞クラス定義を選択し,固有名詞を登録
2. 原言語文を既知語置き換えモデルで翻訳
3. 翻訳結果をBLEUで自動評価、全クラス定義で実施
4. **BLEUスコア最大となる固有名詞クラスと原言語文**を学習コーパス追加

翻訳処理

■**固有名詞の検出, NER情報により既知語**で翻訳

原文	あと二三分で一夜城に着きます
SentencePiece	It takes a couple of minutes to arrive **at the castle overnight**
既知語置換M	We'll get to **Ichiya Castle** in a few minute

実 験

〇 評価
固有名詞検出率・固有名詞正解率・BLEU 評価
〇 実験条件
NER学習手法：　BiLSTM-CRF、GRN、BERT-NER
学習データ：　JParaCrawl
テストデータ：　岐阜タクシーの翻訳社会実証データの2379 文
翻訳エンジン：　NICTのみんなの自動翻訳@ TexTra2（既知語）
SentencePieceモデル（JParaCrawl）

手法	固有名詞検出率	固有名詞正解率	BLEU
人手付与	100.00	100.00	39.26
SentencePiece	0.00	89.86	35.73
BiLSTM-CRF	24.59	89.86	37.40
GRN	45.71	91.30	35.22
BERT-NER	**94.29**	**98.55**	**37.55**

人手付与：人手で単語クラスを付与、既知語モデルで翻訳
SentencePiece：単語登録せずSentencePieceモデルで翻訳