P7-14

クラウドソーシングによる Web サイトマイニングを用いた翻訳モデルの即時領域適応

森下 睦[1, 2], 鈴木 潤[2], 永田 昌明[1]

[1]NTT コミュニケーション科学基礎研究所 , [2] 東北大学

概要

NMT では、十分に学習できていない分野は翻訳品質が低い
→ 特定領域の翻訳精度を高めたい (翻訳モデルの領域適応)
→ 特に COVID-19 のような緊急事態発生時に、
特定領域に特化した機械翻訳モデルを迅速に提供したい

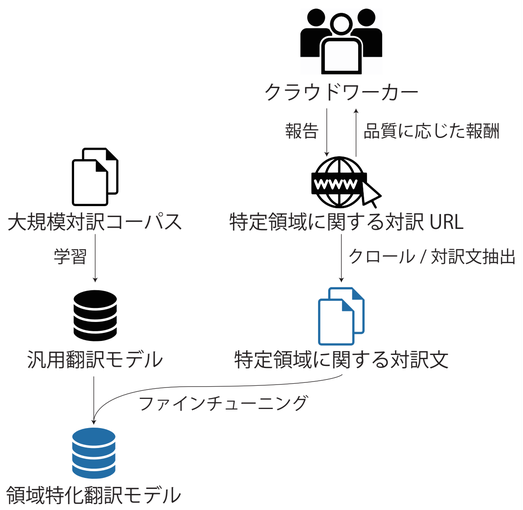
領域適応のためには、特定領域の対訳文が必要
→ クラウドワーカーと協力して
Web から効率よく迅速に特定領域の対訳文を収集する手法を検討

特定領域対訳文収集

クラウドワーカーと協力

特定領域に関する対訳文が存在する URL 対の報告を依頼
→ クラウドワーカーはこれまでの経験から
特定領域の対訳文が存在する
Web サイトを探せるはず

既存手法: CommonCrawl から
日英文が一定以上存在する
Web サイトを見つけクロール
→ CommonCrawl に含まれていない
Web サイトが多数存在
→ 特定領域に特化した対訳文は収集不可



クラウドワーカーへの報酬は品質によって変動

下記スコアを用いた報酬設定により、正しい対訳かつ適応領域に近い Web サイトを探し出すことを期待

対訳品質スコア: 得られた対訳文が正しい対訳か
→ 多言語文埋め込み LASER を用いて対訳の文類似度を計算

領域類似度スコア: 得られた対訳文が適応領域に類似しているか
→ LASER を用いて Dev セットとの文書類似度を計算

実験

実験設定

言語対: 英日翻訳
収集対象領域: COVID-19
テストセット: TICO-19 [1] COVID-19 に関する文を集めたテストセット
(英文を人手で日本語訳し、日英テストセットを作成)
適応前汎用翻訳モデル: Transformer Big モデル , JParaCrawl v2.0 (1000 万文) を学習
クラウドソーシング: 10 人募集 , 5 日間かけて対訳 URL を収集
比較手法: Moore-Lewis [2] (JParaCrawl から適応領域に近い対訳文を自動抽出)

[1] https://tico-19.github.io/
[2] Intelligent Selection of Language Model Training Data, Robert C. Moore and William Lewis, ACL 2010

収集結果

収集対訳 URL 対: 504 件
収集対訳文数: 6,772 文
総報酬額: 31,079 円

	Dev	Test
JParaCrawlのみ	25.9	33.0
Moore-Lewisを用いた領域適応	25.3	33.1
提案法を用いた領域適応	**27.3**	**34.5**

領域適応後の翻訳精度

提案法により収集された対訳文を用いて Fine-tuning
→ 汎用翻訳モデルと比較して 1.5 ポイントの BLEU 向上
→ 既存対訳コーパスから類似対訳文を選ぶ手法 (Moore-Lewis) では、
翻訳精度の改善がみられない

領域別翻訳精度

TICO-19 内の領域別翻訳精度
→ ほぼ全ての領域において翻訳精度が改善
→ 特にニュースで大幅改善 (+4.0)
→ ニュース関連の文は Web から収集しやすい？

	JParaCrawlのみ	領域適応モデル
医療会話	15.3	**16.7**
医学論文	37.8	**38.8**
ニュース	29.4	**33.4**
Wikipedia	30.9	**32.8**
告知文	**26.4**	24.4

収集日数と翻訳精度

収集 2 日目には、大幅に精度改善
→ 緊急事態発生時など領域特化翻訳モデルが
突然必要になった際にも、本手法により
迅速に領域適応可能に