P3-10 教師なし機械翻訳に基づく話し言葉翻訳へのドメイン適応の検討

福田 りょう　須藤 克仁　中村 哲
奈良先端科学技術大学院大学

研究内容

高品質な話し言葉機械翻訳を目的とした
擬似話し言葉によるドメイン適応学習の検討

はじめに | 話し言葉機械翻訳の難しさ

▲ データの不足

- 学習に必要な「対訳データ」の多くが書き言葉
- 話し言葉のデータ作成はコスト大 (時間・金銭)

▲ 書き言葉との隔たり

書き言葉: 文法的

本日午前10時より会議室にてミーティングを行う。

話し言葉: 非文法的 (言い淀み, 品詞の脱落, 区切り…)

あ今日10時からミーティングやるから 会議室で

書き言葉との差異が, 学習データの不足を補う
ドメイン適応学習の障壁になっている

手法 | 擬似話し言葉によるドメイン適応

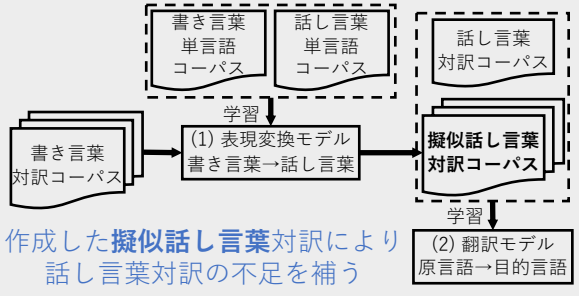
(1) 言語表現変換モデル

(1-1) 書き言葉から話し言葉への言語内翻訳を学習

(1-2) 書き言葉対訳の原言語を擬似話し言葉に変換

(2) 話し言葉翻訳モデル

(2-1) (1-2)で生成した擬似話し言葉を用いて学習



作成した**擬似話し言葉**対訳により
話し言葉対訳の不足を補う

実験(1) | 擬似話し言葉生成

実験設定

システム: Unsupervised MT [Lample et al., 2018]

- 折り返し翻訳による擬似的な教師あり学習
- モデル: Transformer (encoder, decoder層を共有)
  - 次元: 埋め込み層, 隠れ層=512, Feed Forward層=2048
  - サブワード化: BPE (共有語彙16,000)
- 学習データ:
  - 書き言葉データは, 論文抄録対訳コーパス(ASPEC),
    話し言葉データは, 日本語話し言葉コーパス(CSJ),
    日本語日常会話コーパス(CEJC), NAIST授業アーカイブ

単言語データ		文数
書き言葉	ASPEC-JE (日本語)	1,003,602
話し言葉	CSJ	134,477
	CEJC	128,668
	NAIST授業アーカイブ	22,251

教師なし機械翻訳による擬似話し言葉文生成例

書き言葉 (ASPEC)	擬似話し言葉 (CSJ-like ASPEC)
代替フロン中には可燃性のものがあるので注意が必要である。	代替フロン中には、可燃性のものがあるので注意が必要であるということが言えます。
3) 消化管内pH変化	三番に消化管内pH変化です。
超伝導トンネル接合 (STJ) を用いた標題検出器を開発した。	で超伝導トンネル接合ですね、STJを用いた標題検出器を開発しました。
ゴーグル機能として使用するだけでなく、シースルー機能を持たせた。	でゴーグル機能として使用するだけでなくて、ルーシース機能を持たせます。

(段落番号や括弧の除去, 間投詞, 語尾変化, 語の崩壊, 時制変化)

話し言葉データ	折り返し翻訳 BLEU	折り返し翻訳 perplexity
CSJ	**80.98**	**1.617**
CEJC	15.14	15.98
NAIST授業アーカイブ	17.02	20.54
CSJ+CEJC+授業アーカイブ	14.54	17.46

実験結果

CSJを用いた学習が最も高い評価を得た. このモデルを用いてASPECを擬似話し言葉に変換した (**CSJ-like ASPEC**).

生成した擬似話し言葉から3-gram言語モデルを構築し, 話し言葉に対するパープレキシティを測定

- ASPECによる言語モデル (1210.7)と比べ約850減少 (360.7)
- 未知語も約1万減少 (47,757→37,561)

実験(2) | 話し言葉の翻訳学習

2手法のドメイン適応学習による話し言葉翻訳の学習

- Multi-domain学習 (&): ドメイン外データとドメイン外データを混合し学習
- Fine-tuning (→): ドメイン外データで事前学習後, ドメイン内データで追加学習

実験設定

システム: OpenNMT-py [Klein et al., 2017]

- オープンソースのNMTツールキット
- モデル: 6層Transformer
  - 次元: 埋め込み層, 隠れ層=512, Feed Forward層=2048
  - サブワード化: BPE (共有語彙16,000)

対訳データ		対訳数
ドメイン内	NAIST授業アーカイブ	7,031
ドメイン外	ASPEC-JE	1,003,602
	CSJ-like ASPEC	1,003,602

実験結果と考察

- (1)(2)ドメイン適応無しの学習では, 擬似話し言葉を使用することで精度が低下. コーパスの品質の差が原因として考えられる.
- (3)(4)Multi-domain学習では, 書き言葉に対し+7.15ポイント, 話し言葉に対し+1.54ポイント向上. ドメイン外データとして擬似話し言葉を用いることで, ドメイン間距離が近くなり学習が容易になった可能性がある.
- (5)(6)Fine-tuningでは書き言葉に対して-3.03低下. 話し言葉に対してはほぼ同等の結果であった. Fine-tuningはドメイン内データに過剰適合するため, ドメイン外データの品質に大きく影響を受けない.
- 2種類のドメイン適応の組み合わせ学習や, 2段階に渡るFine-tuningなども検討したが, これらで擬似話し言葉が有意に上回る結果は見られなかった.

日英機械翻訳モデルの書き言葉と話し言葉に対するBLEU

適応手法	ASPEC	授業アーカイブ
(1) ASPEC	**27.52**	6.16
(2) CSJ-like ASPEC	23.86	5.58
(3) ASPEC&授業アーカイブ	17.13	6.61
(4) CSJ-like ASPEC&授業アーカイブ	24.28	8.15
(5) ASPEC→授業アーカイブ	23.99	12.71
(6) CSJ-like ASPEC→授業アーカイブ	20.93	**12.81**

※ A&B: AとBのMulti-domain学習. A→B: AからBへのFine-tuning

今後の課題

- 擬似話し言葉データは, Multi-domain学習においては有効性を見せたが, Fine-tuningによる学習においては有意差が見られなかった. フィラーの有無や文体の違いは, ドメイン適応学習, 特にFine-tuningの効果を著しく低下させる特徴では無いと考えられる.
- 語順や文長など, より大胆な言い換えを生成し, 翻訳精度に及ぼす影響を調査する必要がある.