P7-7

Positional Encodingへの摂動付与による長さ制御を用いた非自己回帰型機械翻訳のための知識蒸留

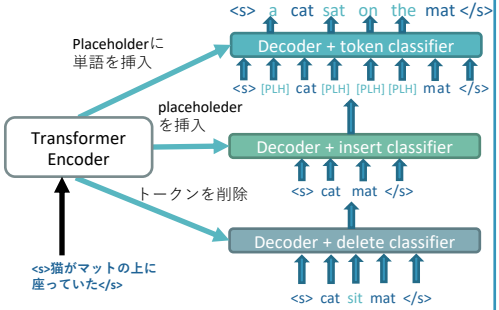
岡 佑依，須藤 克仁，中村 哲（NAIST）

研究概要

▷**問題** 非自己回帰型機械翻訳(NAT)モデルは高速だが翻訳精度が低く，知識蒸留を必要とする
▷**手法** 摂動付きLDPEを知識蒸留に用いるTransformer, Levenshtein Transformerの両方に適用する
▷**結論** 英日,独英翻訳に提案手法は有効であり，参照訳長を制約とした時大幅に改善

関連研究

▷ Levenshtein Transformer [Gu et al. , 2019]



エンコーダ,全てのデコーダはposition embeddingを使う
他のNATモデルと同様知識蒸留を用いる

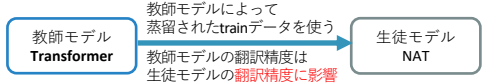
▷ 摂動付きLDPE (Perturbed LDPE) [Oka et al., 2020]

$$perLDPE_{(pos,2i)} = \sin(\frac{len - pos + per}{10000^{2i/d}})$$

$$perLDPE_{(pos,2i+1)} = \cos(\frac{len - pos + per}{10000^{2i/d}})$$

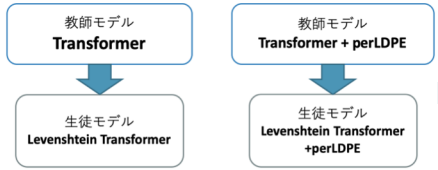
学習時，perは一様分布に基づいたランダム整数値を付与
生成時，per=0（通常のLDPEと同じ）
参照訳長を入力した時翻訳精度が大幅に向上
デコーダにのみ適用，エンコーダは通常のPEを用いる

▷ 文単位の知識蒸留とNAT [Kim et al., 2016 and Zhou et al., 2020]



提案手法

Levenshtein Transformer　提案手法



▷ Transformer + perLDPE (教師モデル)
- 知識蒸留に用いるTransformerにperLDPEを適用
- Trainデータの参照訳長を生成時perLDPEへ入力

→教師モデルの翻訳精度を改善することで生徒モデルの翻訳精度の改善を期待

▷ Levenshtein Transformer + perLDPE (生徒モデル)
- デコーダ(insertionのみ)にperLDPEを適用
- perの摂動範囲は正の値のみ与える

→perLDPEによって出力が長くなることを期待

実験

▷実験条件 コーパス: ASPEC(英日), WMT14(英独,独英)　実装 : fairseq

▷教師モデルの比較 (ベースライン：通常のTransformer)

◆ **perLDPEを用いたTransformer** - 学習時のperLDPEの摂動範囲[-4,+4]，翻訳時は参照訳長を長さ制約

▷生徒モデルの比較 (ベースライン：通常のTransformerを教師モデルとしたLevenshtein Transformer)

◆ **MaskT** – Mask-Predict 目的言語文の一部をマスクして予測するNATモデル [Ghazvininejadet al., 2019]

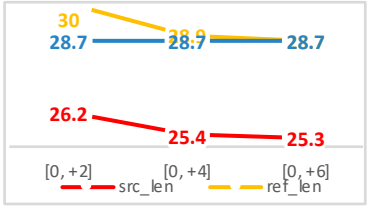
◆ **LevT** – Levenshtein Transformer

◆ **LevT+perLDPE** – 学習時のperLDPEの摂動範囲[0,+2]，翻訳時の長さ制約は英日ではBERTによる予測長，英独/独英では原言語文長

[生徒モデルの比較] モデル	英日 BLEU	英日 LR	英独 BLEU	英独 LR	独英 BLEU	独英 LR
Transformer	37.1	0.948	31.0	0.960	33.0	0.908
教師モデル：通常のTransformer						
MaskT	31.0	0.928	25.9	0.975	28.8	0.880
LevT	34.0	0.912	**28.7**	0.905	27.4	0.838
LevT + perLDPE	33.2	0.897	26.2	0.989	29.4	0.959
LevT + perLDPE	34.2	0.951	30.0	0.997	32.6	0.954
教師モデル：perLDPEを用いたTransformer(提案手法)						
MaskT	**31.3**	0.943	25.9	0.955	28.3	0.884
LevT	**34.3**	0.900	27.4	0.919	28.0	0.839
LevT + perLDPE	34.0	0.918	26.3	0.928	**29.5**	0.951
LevT + perLDPE	34.5	0.988	30.0	0.934	32.7	0.946

表：生徒モデルのBLEUとLR一覧.
英日・独英ではBLEUの改善があったが英独ではなかった.
長さ制約が参照訳の時全てにおいて改善する

図: 生徒モデルの摂動範囲毎のBLEU値の変化 (WMT14英独). 大きくしてもBLEUは改善しない



[教師モデルの比較] モデル	英日 BLEU	英独 BLEU	独英 BLEU
Transformer	32.4	30.1	32.9
perLDPEを用いたTransformer(提案手法)	32.5	31.1	34.9

表: 教師モデルのtrainセットのBLEU. 全ての翻訳において提案手法は有意であった.

青色:ベースライン,
赤色:提案手法の長さ制約, 黄色:参照訳長の長さ制約