P1-20

Positional Encoding 出力長制御を用いた英日ニューラル機械翻訳の検討

岡 佑依，帖佐 克己，須藤 克仁，中村 哲（NAIST）
Mail : oka.yui.ov2@is.naist.jp

Abstract

[動機] ニューラル機械翻訳における，**訳抜け問題，重複訳問題を翻訳モデルに出力長を与えることで解決**
[手法] 要約文生成で適用された位置エンコーディングに基づく出力長制御を翻訳タスクに適用
[結果] 学習時に与えるトークン長に幅を与える方がBLEUスコアが上がった
訳抜け・重複訳が解決した文と，逆に発生してしまった文が見受けられた

提案手法

[先行研究]
位置エンコーディングを用いた出力長制御 [Takase+, NAACL-HLT 2019]

・**LDPE** – 文の終端までの差に応じた位置
・**LRPE** – 文の終端までの比率に応じた位置
この二つの式をTransformerモデルに適用
文字数単位での出力長制御が可能

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

[本研究]
・LRPE，LDPEを翻訳に適用
・本研究ではLRPE,LDPEにトークン長を与える
・入力文からトークン長を予測し，正しいトークン長を与えると訳抜け，重複訳が解決するのではないか？
・このアプローチの有効性を検証するため，正解トークン長に基づく出力長制御を行った
・学習時に出力長制約に幅(±0-2)を持たせた長さを与えた場合も検証した

実験結果

[実験条件]
コーパス：ASPEC(1,000,000/1790/1812)　トークナイズ：SentencePiece　ボキャブラリサイズ：16,000
ハイパーパラメータ：エンコーダ・デコーダ層数２，その他は(Vaswani et al., 2017)

[実験結果] **重複訳が発生した例**

ref	3相相関イメージセンサー(three－phase correlation image sensor:3PCIS)を使った実時間の位相スタンプ・レンジファインダー(phase－stamp range finder:PSRF)を提案する。
baseline	3相相関イメージセンサ(3PCIS)を用いたリアルタイム位相スタンプ距離計(RT-PSRF)を提案した 【25】
LDPE	三次元位相相関イメージセンサ(3PCIS)を用いた実時間位相スタンプレンジファインダ(RT-PSRF）を提案し,その動作作理と特徴,特徴,特徴,特徴,特徴などを紹介した。【44】
LRPE	3相相関イメージセンサ(3PCIS)を用いた実時間位相スタンプ距離計(RT-PSRF)を提案し,その原理,原理,原理,原理,原理,および原理,原理,特性について述べた。【45】
LRPE ±0-2	3相相関イメージセンサ(3PCIS)を用いた実時間位相スタンプレンジファインダ(RT-PSRF)を提案し,その有効性を検証するために,3相相関イメージセンサ(3PCIS)を用いた。【45】

LDPE,LRPEと比べ，出力トークン数がほぼ同じだがLRPE ±0-2は重複訳の回数が少ない

モデル	学習時 len	BLEU	分散
ベースライン		41.21	22.312
LRPE	tgt_len	39.79	0.167
	tgt_len ± 0~2	41.56	3.986
LDPE	tgt_len	39.32	0.001
	tgt_len ± 0~2	40.30	2.236

翻訳時ではなく，学習時に与える長さに”揺らぎ”を与えるとBLEUが上がった

出力長制御の精度を検証

$$\frac{1}{n}\sum_{i=1}^{n}|ref_i - mt_i|^2$$

分散が小さいことから出力長制御ができているとわかる．
LDPEの方が精度が良い

[考察]
ベースラインに比べ，提案手法が正しいトークン長を得た時，訳抜け・重複訳が解決している例は多く見受けられた
原言語文に含まれていないフレーズを翻訳するとき，ベースラインは短く出力することで翻訳文の精度を維持するが，提案手法では長さがすでに与えられているため，重複訳をするという問題も発生したと考えられる．