P4-24　汎用的な文の分散表現を用いた文単位の機械翻訳自動評価

首都大学東京　shimanaka-hiroki@ed.tmu.ac.jp　嶋中宏希　梶原智之　小町守

1. 研究の背景と概要

- WMTのMetricsタスクで優秀な成績を収めたほとんどの手法が文字N-gramや単語N-gramといった，表層に基づく素性を利用しており限定的な情報しか扱えていない．
- そこで本研究では，汎用的な文の分散表現を利用し，文字N-gramや単語N-gramなどの表層に基づく情報を超えた広範な情報を考慮した機械翻訳評価手法を提案する．
- 実験の結果，我々の提案手法は**文の分散表現のみを素性として用いた回帰モデル**で**最高性能を達成**した．

例:　翻訳文: This is not a major issue.
　　参照文: It is nothing major.

人手評価値:	0.892	(32/560)
Blend:	-0.0734	(423/560)
提案手法:	**0.554**	**(60/560)**

2. 従来手法

- ReVal [Gupta et al., 2015]
  - Tree-LSTMで文単位の類似度値を用いて学習する手法．
  - 少ないデータで文の分散表現の学習をするためあまり良い結果を残せていない．
- Blend [Ma et al., 2017]
  - SVR (RBF カーネル) で人手評価値を用いて学習する手法．
  - 素性: Asiyaのデフォルト素性（字句ベース），文字N-gram，文字単位の編集距離

3. 提案手法

本研究では，事前学習された汎用的な文の分散表現 (Skip-Thought / InferSent)を用いてSVR (RBF カーネル) で人手評価値の学習を行い機械翻訳の自動評価を行う．

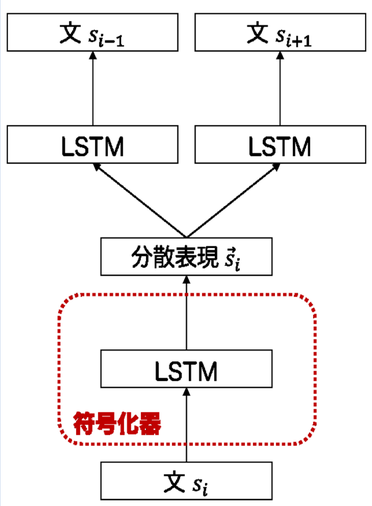

図 1: Skip-Thoughtの概要

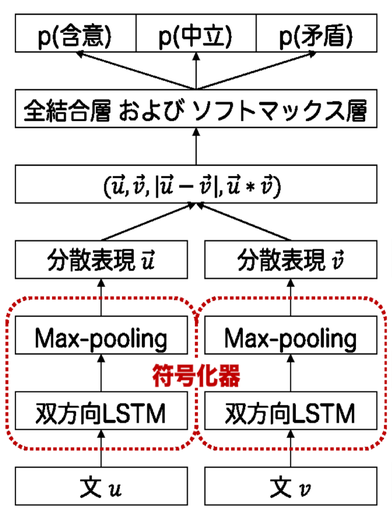

図 2: InferSentの概要

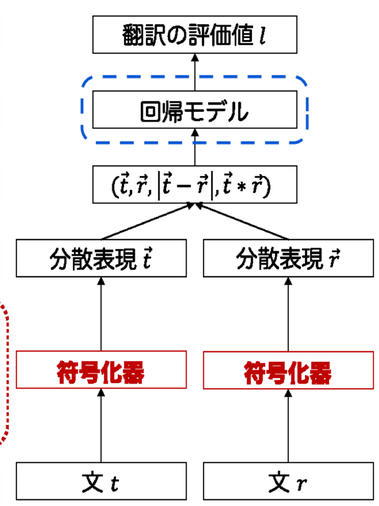

図 3:本研究の概要

4. 実験設定

◆ 汎用的な文の分散表現

本研究では既存の学習済みのものを用いた．

- Skip-Thought [Kiros et al., 2015]
  学習データ: Aligning books and movies
  文の分散表現の次元: 4,800次元
- InferSent [Conneau et al., 2017]
  学習データ: The Stanford Natural Language Inference (SNLI) Corpus
  文の分散表現の次元: 4,096次元

◆ 人手評価値の学習データセット

表 1: WMTの人手評価値付き対訳文の文対数

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
WMT-2015	500	500	500	-	500	-
WMT-2016	560	560	560	560	560	560

◆ 機械翻訳自動評価のための回帰モデル

- scikit-learn の SVR (RBF カーネル)

$C \in \{0.01,0.1,1.0,10\}$
$\varepsilon \in \{0.01,0.1,1.0,10\}$
$\gamma \in \{0.01, 0.1, 1.0, 10\}$

グリッドサーチと10 分割交差検定を行った．

5. 実験結果

表 2: ピアソンの相関係数 (newstest2016)

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Avg.
SentBLEU	0.557	0.448	0.484	0.499	0.502	0.532	0.504
Blend [Ma et al., 2017]	0.709	0.601	0.584	0.636	0.633	**0.675**	0.640
$\text{DPMF}_{\text{comb}}$ [Yu et al., 2015]	**0.713**	0.584	0.598	0.627	0.615	0.663	0.633
ReVal [Bojar et al., 2016]	0.577	0.528	0.471	0.547	0.528	0.531	0.530
Skip-Thought	0.665	0.571	0.609	**0.677**	0.608	0.599	0.622
InferSent	0.679	0.604	0.617	0.640	0.644	0.630	0.636
InferSent + Skip-Thought	0.686	**0.611**	**0.633**	0.660	**0.649**	0.646	**0.648**

6. 考察

表 3: 全言語対についての人手評価値に基づく上位20%に対する分析結果（合計 672文）

	Blend でのみ正しく評価（全 70 文）	提案手法でのみ正しく評価（全 88 文）
表層の一致率が低い	26	42
未知語を含む (かつ文長が短い)	26 (17)	26 (2)
その他	24	31

- 提案手法でのみ正しく評価できている翻訳文が多いことから，参照文と意味が似ている翻訳文において提案手法のほうが良い結果を示せていると考えられる．
- 提案手法では表層の一致率が低い翻訳文に対しても正しい評価ができている翻訳文が多く，字句ベースの評価手法では捉えきれなかった文の情報を捉えれていると考えられる．
- 提案手法では，参照文と意味が似ている翻訳文において，文長の長い (15単語より長い) 文においては未知語が存在していても正しい評価が行えている文が多く，未知語による影響は小さいと考えられる．