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

嶋中 宏希　　梶原 智之　　小町 守
首都大学東京
{shimanaka-hiroki, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

本研究では，参照文を用いた文単位での機械翻訳自動評価手法について述べる．人手評価との相関が高い文単位の評価ができることにより機械翻訳システムの細かい改善が可能になる．

機械翻訳に関する国際会議 WMT では，2008 年から自動評価手法の性能を競う Metrics タスクが開催されており，これまでに多くの機械翻訳自動評価手法が提案されてきた．しかし，現在のデファクトスタンダードである BLEU [1] をはじめ，WMT-2017 の Metrics タスク [2] で優秀な成績を収めた Blend [3], MEANT 2.0 [4], CHRF++ [5] など，ほとんどの機械翻訳自動評価手法が文字 N-gram または単語 N-gram に基づく素性を利用しており，文単位での評価にとっては限定的な情報しか扱えていない．

そこで本研究では，文単位での表現学習によって獲得した文の分散表現を利用し，単語 N-gram 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．WMT-2016 のデータセットを用いた実験の結果，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．

2 関連研究

Blend[1] [3] は，WMT-2016 の Metrics タスク [6] で最高性能を達成した $\text{DPMF}_{\text{comb}}$ [7] と同様に様々な評価手法のスコアを素性として用いる SVR（RBF カーネル）モデルであり，WMT-2017 の Metrics タスク [2] で最高性能を達成している．Blend は機械翻訳の自動評価用ツールキット Asiya[2] [8] のデフォルトの字句ベースの評価手法のスコア 25 素性に加え，文字 N-gram と置換木に基づく線形モデルである BEER [9]，文字単位の編集距離に基づいて評価する CharacTER [10]，翻訳文と参照文の構文的な類似性を評価する DPMF [11] および翻訳文の流暢性を評価する ENTF [12] を素性として利用している．

$\text{DPMF}_{\text{comb}}$ は同じ原文についての複数の翻訳文に対して，原文とそれぞれの翻訳文を比較することにより付けられた，人手による**相対的な品質評価**（RR : Relative Ranking）のデータを用いて，Ranking SVM で学習を行っているが，Blend はすべての翻訳文について翻訳文と参照文を比較することにより付けられた，人手による**絶対的な品質評価**（DA : Direct Assessment）のデータを用いて SVR で学習を行っている．WMT-2016 の Metrics タスクのデータを用いた実験において，Blend の方が $\text{DPMF}_{\text{comb}}$ よりも良い性能を示している（表 2）．本研究でも，絶対評価を用いた教師あり回帰モデルを提案する．

ReVal[3] [13] は，相対評価を擬似的な文の類似度スコアのラベル付きデータに変換し，生成された擬似的な文の類似度スコアのデータと別ドメインである，SICK[4] の類似度スコアが付いた文単位のラベル付きデータを組み合わせて学習を行う機械翻訳評価手法である．この手法では，Tree-LSTM [14] を用いて文の広範な情報を考慮している．ただし，この手法で用いられた学習データは約 21,000 文と少なく，Tree-LSTM の学習は不安定であり正確な学習が難しい（表 2）．本研究でも LSTM を用いた文の表現学習を行うが，我々は他のタスクで大規模なデータから訓練され，汎用的であると示されている文の分散表現を本タスクに適用する．そのため，文の表現学習に用いるデータが少ないという問題を避けることができる．

3 文の分散表現を用いた教師あり学習に基づく機械翻訳自動評価手法

本研究では，事前学習された汎用的であると示されている文の分散表現を用いて，機械翻訳の自動評価を

[1] http://github.com/qingsongma/blend
[2] http://asiya.lsi.upc.edu/
[3] https://github.com/rohitguptacs/ReVal
[4] http://clic.cimec.unitn.it/composes/sick.html

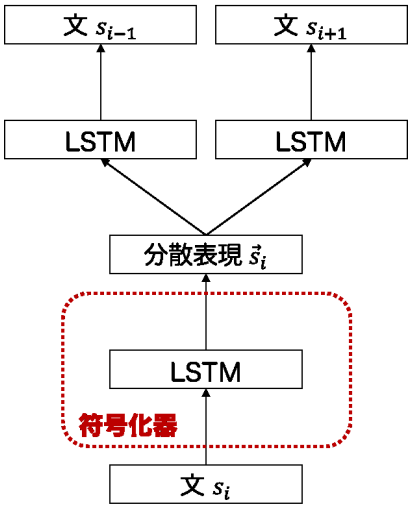

図 1: Skip-Thought の表現学習

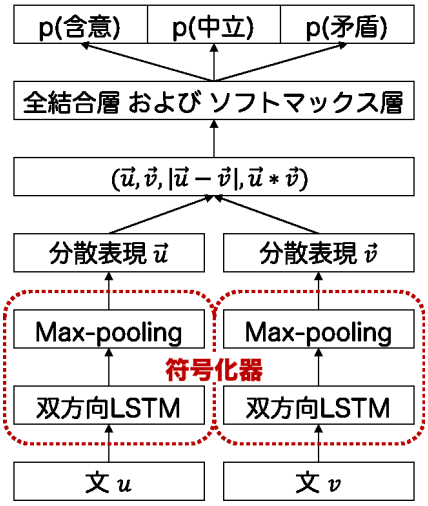

図 2: InferSent の表現学習

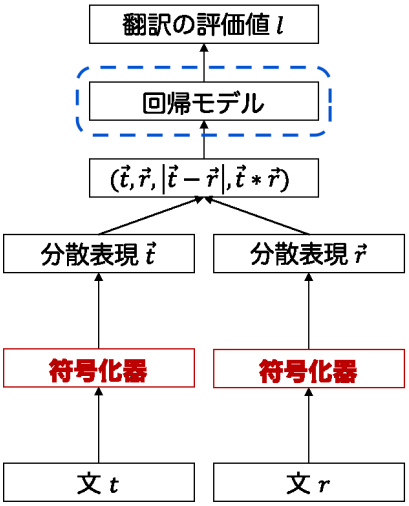

図 3: 本研究の機械翻訳自動評価

行う．まず 3.1 節では，本研究で使用する 3 種類の文の分散表現について説明する．続いて 3.2 節では，機械翻訳自動評価のための回帰モデルおよび素性抽出について述べる．

3.1 汎用的な文の分散表現

Skip-Thought[5] [15] は，連続する 3 文 s_{i-1}, s_i, s_{i+1} を用いる教師なし学習に基づく手法である．文 s_i を符号化し，その文の分散表現 $\vec{s_i}$ から，前後の文の分散表現 $\vec{s_{i-1}}$ および $\vec{s_{i+1}}$ を予測するというエンコーダデコーダモデル（図 1）を学習することによって，符号化器を用いた文の分散表現が可能になる．Skip-Thought は，特に文書分類タスクへの応用で高い性能を発揮する．

InferSent[6] [16] は，含意関係認識のための SNLI データセット[7] [17] 上で Max-pooling を用いた双方向 LSTM ネットワークを訓練し，教師ありで文の表現学習を行う手法である．文 u と v をそれぞれ符号化し，それらの文の分散表現 $\vec{u}$ と $\vec{v}$ から素性を生成し，その素性を用いて含意関係認識のデータセットを教師データとして文の表現学習を行うモデル（図 2）である．含意関係認識は所与の文対の関係を含意／矛盾／中立に 3 値分類するタスクであり，意味の違いに敏感な文の分散表現が得られると期待できる．そのため InferSent は，文書分類と意味的類似度の両応用タスクで安定して高い性能を発揮する．

3.2 機械翻訳自動評価のための回帰モデル

本研究では，文単位での機械翻訳の自動評価を行う．この問題は，翻訳文 t および参照文 r から実数で表現される翻訳の評価値 l を推定する回帰問題として扱うことができる．共通の符号化器を用いて d 次元に符号化された翻訳文の分散表現 $\vec{t}$ および参照文の分散表現 $\vec{r}$ を用いて，InferSent に倣って以下の 3 つの方法で翻訳文と参照文の関係を素性に組み込む．

- 連結：$(\vec{t}, \vec{r})$
- 要素積：$\vec{t} * \vec{r}$
- 要素差：$|\vec{t} - \vec{r}|$

回帰モデルには，これらの 3 種類の素性を連結した $4d$ 次元の素性が入力される（図 3）．提案手法では回帰モデルのみの学習を行い文の表現学習は行わない．

また，複数の種類の符号化器を用いた実験も行った．この実験は，それぞれの符号化器について生成した前述の $4d$ 次元の素性を，更に符号化器の種類数だけ連結した素性を回帰モデルに入力したものである．

4 回帰モデルを用いた英語方向における文単位の機械翻訳評価実験

WMT の Metrics タスクの英語方向のデータセットを用いて提案手法の性能を検証する．

[5] https://github.com/ryankiros/skip-thoughts
[6] https://github.com/facebookresearch/InferSent
[7] https://nlp.stanford.edu/projects/snli/

[8] en : English, cs : Czech, de : German, fi : Finnish ro : Romanian, ru : Russian, tr : Turkish

表 1: WMT-2015 [18] と WMT-2016 [6] の英語方向の各言語対における絶対評価のデータ数[8]

	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

表 2: 英語方向の各言語対におけるピアソンの相関係数を用いた各手法の評価結果 (newstest2016)

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Avg.
SentBLEU [6]	0.557	0.448	0.484	0.499	0.502	0.532	0.504
Blend[9] [3]	0.709	0.601	0.584	0.636	0.633	**0.675**	0.640
$\text{DPMF}_{\text{comb}}$ [6]	**0.713**	0.584	0.598	0.627	0.615	0.663	0.633
ReVal [6]	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**

4.1 実験設定

データセット 評価には WMT-2016 の Metrics タスク [6] の newstest2016 の英語方向のデータセットを用いて行った．なお，学習には WMT-2015 [18] と WMT-2016 の Metrics タスクの newstest2015 と newstest2016 のデータセットから各言語対に対して，評価データ以外のデータ 4,800 文を学習データとして用いた（表 1）．これらのデータセットは，すべての翻訳文について，翻訳文と参照文を比較することにより付けられた，絶対評価が付けられている．

モデル 提案手法において，回帰モデルの学習では 3.2 節で述べた素性に対し，scikit-learn[10] の SVR（RBF カーネル）モデルを用いて行った．SVR モデルの各パラメータについては $C \in \{0.01, 0.1, 1.0, 10\}$, $\epsilon \in \{0.01, 0.1, 1.0, 10\}$, $\gamma \in \{0.01, 0.1, 1.0, 10\}$ でグリッドサーチと 10 分割交差検定を行うことにより決定した．

比較手法は，2 節で示した Blend [3], $\text{DPMF}_{\text{comb}}$ [7], ReVal [13] の 3 つである． Blend と $\text{DPMF}_{\text{comb}}$ は，それぞれ WMT-2017 の Metrics タスク [2] と WMT-2016 の Metrics タスクで最高性能を示した機械翻訳自動評価手法である．各手法の評価にはピアソンの相関係数を用いて絶対評価との相関を見ることにより比較を行った．

4.2 実験結果

実験結果を表 2 に示す．表 2 より，提案手法の InferSent [16] と Skip-Thought [15] を組み合わせて使ったモデルで， WMT-2016 の Metrics タスクや WMT-2017 の Metrics タスクにおいて上位の結果を残している全ての手法より良い結果を得た．

これらの結果より，汎用的な文分散表現は翻訳における絶対評価を学習することにより，機械翻訳評価手法への適用が可能であり，なおかつ絶対評価との相関が高いということが示された．様々な評価手法のスコアを素性として用い絶対評価の学習を行っている Blend を超える良い結果を得られたことから，汎用的な文分散表現には様々な評価手法を組み合わせることにより捉えられる文の情報を含んでいると考えられる．

また，ReVal では文単位で絶対評価との高い相関を得られてなかったが，提案手法では絶対評価との高い相関が得られた．このことから，機械翻訳における相対評価や文の類似度スコアなどの少ないデータを用いて学習を行った文分散表現を使った評価手法より，事前に十分なデータで学習を行った文の分散表現で少ない絶対評価を学習する評価手法のほうが機械翻訳の評価には有効であること考えられる．

4.3 分析

Blend[11] [3] の再実装を行い，Blend と提案手法の評価結果の比較を行った．[12]それぞれの言語対で高い人

[9] この結果は論文から引用した値である．
[10] http://scikit-learn.org/stable/
[11] http://github.com/qingsongma/blend
[12] 再実装による Blend の全言語対の平均スコアは 0.636 となり論文で報告されている値より少し低い値になっているが，以下の議論に影響はないと判断した．

手評価のスコアが付けられた，つまり，翻訳文の意味が参照文の意味に近い上位 56 文に対して，Blend では正しく評価できているが提案手法では正しく評価できていない文，提案手法では正しく評価できているが Blend では正しく評価できていない文について分析を行った．Blend でのみ正しく評価できている文の全言語対の合計が 46 文，提案手法でのみ正しく評価できている文の全言語対の合計が 42 文となった．

Blend では正しく評価できている全 46 文のうち，表層の一致率が高い文が 43 文，表層の一致率が低い文が 3 文という結果になった．Blend が様々な表層に基づく評価手法のスコアを用いて学習を行っていることから，この結果は明白であると考えられる．また，提案手法が正しく評価できていなかった理由に，意味は全く同じだが表層が違う語やフレーズが含まれる文が，17 文も存在した．これは汎用的な文の分散表現に既存の学習済みのものを用いたためである，データの表層を揃えて文の分散表現を学習し直すことにより、提案手法のさらなる正確な評価が可能になると考えられる．

提案手法では正しく評価できている全 42 文のうち，表層の一致率が高い文が 27 文，表層の一致率が低い文が 15 文という結果になった．このことから，提案手法は参照文との類似度が高い翻訳文に対して，Blend では捉えきれない表層以外の文の特徴も捉えられていると考えられる．

5 おわりに

本研究では事前に学習を行った文の分散表現から素性を生成し，絶対評価の学習を行うことにより機械翻訳自動評価手法への適用を試みた．実験の結果，汎用的な文分散表現を用いて絶対評価を教師データとして学習することにより，少ない絶対評価のデータでも絶対評価との非常に高い相関が得られることが示された．

今回は文単位の英語方向についての実験しか行えていない．そこで，提案手法がシステム単位での機械翻訳自動評価にも有効であるかという実験も行いたい．また，様々な言語の文の分散表現の学習が可能な手法と様々な言語対についての絶対評価が存在すれば提案手法は有効なので，その両方が存在する英語以外の言語についても実験も行いたいと考えている．

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pp. 311–318, 2002.

[2] Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 Metrics Shared Task. In *Proc. of WMT*, pp. 489–513, 2017.

[3] Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. Blend: a Novel Combined MT Metric Based on Direct Assessment —CASICT-DCU submission to WMT17 Metrics Task. In *Proc. of WMT*, pp. 598–603, 2017.

[4] Chi kiu Lo. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proc. of WMT*, pp. 589–597, 2017.

[5] Maja Popović. CHRF++: words helping character n-grams. In *Proc. of WMT*, pp. 612–618, 2017.

[6] Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 Metrics Shared Task. In *Proc. of WMT*, pp. 199–231, 2016.

[7] Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proc. of WMT*, pp. 417–421, 2015.

[8] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *The Prague Bulletin of Mathematical Linguistics*, No. 94, pp. 77–86, 2010.

[9] Miloš Stanojević and Khalil Sima'an. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proc. of WMT*, pp. 396–401, 2015.

[10] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. Character: Translation edit rate on character level. In *Proc. of WMT*, 2016.

[11] Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. An Automatic Machine Translation Evaluation Metric Based on Dependency Parsing Model. *arXiv preprint arXiv:1508.01996*, 2015.

[12] Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. Improve the Evaluation of Fluency Using Entropy for Machine Translation Evaluation Metrics. *arXiv preprint arXiv:1508.02225*, 2015.

[13] Rohit Gupta, Constantin Orasan, and Josef van Genabith. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proc. of EMNLP*, pp. 1066–1072, 2015.

[14] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proc. of ACL*, pp. 1556–1566, 2015.

[15] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In *Proc. of NIPS*, pp. 3294–3302, 2015.

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proc. of EMNLP*, pp. 670–680, 2017.

[17] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *Proc. of EMNLP*, pp. 632–642, 2015.

[18] Miloš Stanojević, Philipp Koehn, and Ondřej Bojar. Results of the WMT15 Metrics Shared Task. In *Proc. of WMT*, pp. 256–273, 2015.