E-011

機械翻訳のためのベクトル間類似度に基づく参照訳を用いない自動評価法

Reference-free Evaluation Metric based on Similarity between Vectors for Machine Translation

藤﨑晴大† 越前谷博†
Haruto Fujisaki† Hiroshi Echizen'ya†

北海学園大学†
Hokkai-Gakuen University†

１．はじめに

近年,機械翻訳のための参照訳を用いない様々な自動評価法が提案されている[1][2].本研究では,原文とその訳文,そして,人手スコアによる参照訳を用いない新たな自動評価法の提案とその性能評価について述べる.提案手法では,原文の文ベクトルとの間のコサイン類似度が適切な評価スコアとなるような文ベクトルを求める.WMT22(the 2022 conference on machine translation)の自動評価タスクデータ[3]を用いた性能評価実験の結果,提案手法の DA スコアに基づく評価精度においては,従来手法における参照訳を用いない自動評価法の中で上位に位置し,提案手法の有効性が示された.

２．提案手法

２．１ 概要

提案手法では,全結合のニューラルネットワークに基づき学習したベクトル変換モデルを用いて文ベクトルを生成し,原文の文ベクトルとの間のコサイン類似度を求めることで自動スコアを得る.図 1 に提案手法の概要図を示す.

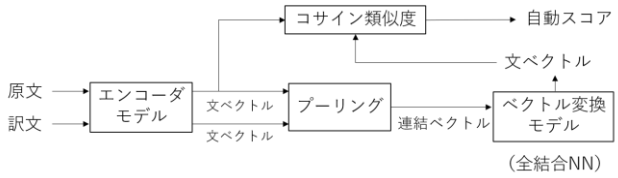

図 1　提案手法の概要図

図1より,原文とその訳文を事前学習済みエンコーダモデルに与え,それぞれ文ベクトルを得る.それら 2 つの文ベクトルをプーリングにより,1 つの文ベクトルに連結し,それをベクトル変換モデルに入力することで,文ベクトルを生成する.生成された文ベクトルは,原文との間の適切な類似度,すなわち,評価スコアが得られるように変換された訳文の文ベクトルに相当する.最後に,変換された文ベクトルと原文の文ベクトルとの間のコサイン類似度を求めることで,自動スコアを得る.

２．２ ベクトル変換モデルの構築

図 1 のベクトル変換モデルは全結合のニューラルネットワークを学習することで構築する.図2にベクトル変換モデルの構築の流れを示す.ベクトル変換モデルの学習の際に用いられる正解ベクトルは原文とその訳文,そして,訳文に対する人手スコアを用いて生成する.損失関数には,正解ベクトルとベクトル変換モデルにより生成された文ベクトルによるCosineEmbeddingLossを用いる.図2のベクトル変換処理では, 原文とその訳文の間のコサイン類似度が人手スコアと等価になるように,訳文の文ベクトルを変更するこ

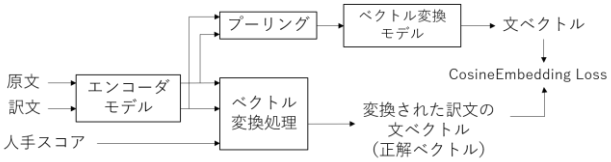

図 2　ベクトル変換モデルの構築

とで正解となる文ベクトルを得る.具体的には原文の文ベクトルを $A(a_1, a_2, \cdots, a_n)$, 訳文の文ベクトルを $B(b_1, b_2, \cdots, b_n)$, また,訳文の文ベクトル B の第 1 要素に加える値を x とすると,以下のコサイン類似度に基づく式(1)より,x を求めることで正解ベクトルが得られる.

$$\frac{a_1 \times (b_1 + x) + \sum_{i=2}^{n}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} {a_i}^2} \times \sqrt{(b_1 + x)^2 + \sum_{i=2}^{n} {b_i}^2}} = 人手スコア \quad (0.0 \sim 1.0) \qquad \cdots(1)$$

この場合,式(1)は x の 2 次方程式となるため,x の値は,2 つの異なる実数解,重解,異なる 2 つの虚数解のいずれかとなる.2 つの異なる実数解が得られた場合には,値の小さな方を x の値として用いる.また,2 つの虚数解が得られた場合には x の値としては 0 を用いる.このように事前学習済みエンコーダモデルより出力される訳文の文ベクトルに対して,変更を加えることにより正解ベクトルを生成する.そして,ベクトル変換モデルより出力される文ベクトルと正解ベクトルを損失関数である CosineEmbeddingLoss に適用することで,ベクトル変換モデルは生成した文ベクトルと原文の文ベクトルとの間の類似度が人手スコアと等価になるように学習される.

３．性能評価実験

３．１ 実験方法

図 2 の提案手法におけるベクトル変換モデルの構築においては,学習データとして,WMT17 から WMT20 の自動評価タスクデータの原文とその訳文の 887,258 ペアと訳文に対する DA スコアを用いて行った.DA スコアは 0～100 のスコアとなっているため,0.0～1.0 のスコアに正規化して正解ベクトルを得た.ベクトル変換モデルの構成は,入力層,2 つの中間層,そして出力層からなる全結合層とした.また,原文とその訳文の文ベクトルを得るためのエンコーダモデルには,xlm-roberta-base[4]を用いた.言語ペアの数は 33 ペアである.さらに本実験では正解ベクトルを得るための人手評価として,WMT20 と WMT21 より提供されている 69,529 の MQM スコアも用いた.MQM スコアに対応する原文とその訳文のペア数も 69,529 である.その際の言語ペアは英語→ドイツ語,英語→ロシア語,中国語→英語である.なお,MQM スコアについても 0.0～1.0 のスコアに正規化して正解ベクトルを得た.このように本研究では,DA スコア用と MQM スコア用のベク

表 1　セグメントレベルの DA スコアのケンドールの順位相関係数

自動評価	cs-en	de-en	ja-en	ru-en	uk-en	zh-en	平均
HWTSC-TLM	0.030	0.011	0.097	0.013	0.001	0.013	0.0275
HWTSC-Teacher-Sim	0.018	0.016	0.098	0.007	0.007	0.001	0.0245
UniTE-src	0.026	0.018	0.087	0.001	0.003	0.007	0.0237
COMETKiwi	0.028	0.011	0.091	0.001	0.004	0.002	0.0228
MS-COMET-QE-22	0.022	0.011	0.088	-0.002	0.003	0.001	0.0205
Cross-QE	0.015	0.011	0.087	0.003	0.001	-0.000	0.0195
REUSE	0.002	0.009	0.091	-0.007	0.000	0.011	0.0177
COMET-QE	0.010	0.020	0.076	-0.005	-0.002	0.003	0.0170
KG-BERTScore	0.010	0.007	0.087	-0.012	0.008	-0.002	0.0163
提案手法	0.016	0.015	0.098	0.003	0.003	0.019	0.0257

表 2　セグメントレベルの MQM スコアのケンドールの順位相関係数

自動評価	en-de	en-ru	zh-en	平均
COMETKiwi	0.290	0.359	0.364	0.338
COMET-QE	0.281	0.341	0.365	0.329
UniTE-src	0.287	0.342	0.343	0.324
Cross-QE	0.263	0.310	0.378	0.317
MS-COMET-QE-22	0.233	0.305	0.287	0.275
MATESE-QE	0.244	0.229	0.337	0.270
HWTSC-Teacher-Sim	0.155	0.143	0.272	0.190
KG-BERTScore	0.129	0.111	0.219	0.153
HWTSC-TLM	0.092	0.121	0.086	0.100
REUSE	0.065	0.078	0.130	0.091
提案手法	0.215	0.227	0.305	0.249

トル変換モデルをそれぞれ構築した.

図 1 の提案手法による自動スコアの生成においては,原文とその訳文の文ベクトルを得るためのエンコーダモデルには,xlm-roberta-base を用いた.また,DA スコア用のベクトル変換モデルを用いる場合の原文とその訳文の言語ペアは WMT22 より提供されているチェコ語→英語(cs-en),ドイツ語→英語(de-en),日本語→英語(ja-en),ロシア語→英語(ru-en),ウクライナ語→英語(uk-en),中国語→英語(zh-en)である.その内訳は,cs-en が 18,824 ペア,de-en が 21,824 ペア,ja-en が 30,120 ペア,ru-en が 24,192 ペア,uk-en が 20,180 ペア,zh-en が 37,500 ペアである.MQM スコア用のベクトル変換モデルを用いる場合も WMT22 より提供されている言語ペアである,en-de,en-ru,zh-en を用いた.その内訳は en-de が 34,629 ペア,en-ru が 34,629 ペア,zh-en が 37,500 ペアである.生成された自動スコアに対しては WMT22 の DA スコアと MQM スコアそれぞれについて人手スコアとの相関係数を用いて評価を行った.

３．２ 実験結果

表 1 に DA スコアにおける提案手法と従来手法のセグメントレベルの相関係数を示す.また,表 2 に MQM スコアにおける提案手法と従来手法のセグメントレベルの相関係数を示す.なお,従来手法は全て参照訳を用いない自動評価である.

表 1 より,DA スコアにおいては,提案手法の平均は 0.0257 となり,他の手法の中でも 2 番目に高い値であった.したがって,提案手法の有効性を示していると考えられる.また,表 2 より,MQM スコアにおいては,提案手法の平均は 0.249 となり,他の手法に対して十分とはいえない.

３．３ 考察

表 1 において,ja-en と zh-en の相関係数が他の手法に比べて最も高い結果となった.その理由としては zh-en は他の言語ペアに比べ,学習データが多く,ベクトル変換モデルの構築に効果的であったと考えられる.ja-en については,zh-en での漢字に対する学習が ja-en での漢字の学習にも有効となり,評価精度が向上したと考えられる.表 2 においては,提案手法の相関係数はいずれの言語ペアにおいても順位は中間に位置しており不十分であった.原因としては DA スコアの学習データ量に対して,MQM スコアの学習データ量は不十分であったと考えられる.

４． まとめ

本研究では,参照訳を用いない新たな自動評価法を提案した.提案手法では,原文の文ベクトルとの間のコサイン類似度が人手スコアと等価になるようにベクトル変換モデルは文ベクトルを生成する.これは原文に対する訳文の文ベクトルを変換することに相当する.性能評価実験の結果,DA スコアにおける評価精度は他の手法に対し,比較的高い精度を示した.したがって,提案手法が参照訳を用いない自動評価において有効であることを示している.

今後は MQM スコアにおける評価精度の向上を図る予定である.

参考文献

[1] R. Rei, A. Farinha, C. Zerva, D. Stigt, C. Stewart, P. Ramos, T. Glushkova, A.Martins, A. Lavie, Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task, In Proc of WMT 2021, pp. 1030-1040 2021

[2] F. Kepler, J. Trénous, M. Treviso, M. Vera, A. Martins, OpenKiwi: An Open Source Framework for Quality Estimation, In Proc of ACL 2019, pp. 117-122 2019

[3] T. Kocmi, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, T. Gowda, Y. Graham, R. Grundkiewicz, B. Haddow, R. Knowles, P. Koehn, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, M. Novák, M. Popel, M. Popović, Findings of the 2022 Conference on Machine Translation (WMT22), In Proc of WMT 2022, pp. 1-45

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov,Unsupervised Cross-lingual Representation Learning at Scale, In Proc of ACL 2020, pp.8440-8451, 2020