柔らかいジャンプ付き編集距離に向けて

亀井 遼平 [1]　横井 祥 [1,2]　仲村 祐希 [1]　渡辺 太郎 [3]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3] 奈良先端科学技術大学院大学
{ryohei.kamei.s4, yuki.nakamura.r1}@dc.tohoku.ac.jp,
{yokoi, kentaro.inui}@tohoku.ac.jp, taro@is.naist.jp

概要

本論文では翻訳システムに対する自動評価指標として新たな指標を提案する．これは単語レベルの編集距離をジャンプ操作の追加と単語埋め込みによる置換コストの緩和によって拡張したものである．前者は語順の入れ替えを，後者は文中の単語の類似度をそれぞれ考慮しており，これらを組み合わせることで質の高い自動評価指標を作れると考えた．実験は WMT19 Metrics Shared Task を用いて行い，上記の拡張が自動評価指標と人手評価の相関向上に寄与することを確認した．

1　はじめに

優れた機械翻訳システムを開発するには質の高い自動評価指標が必要であり [1]，以前から BLEU [2] や chrF [3] などが頻繁に用いられてきた．自動評価指標の質を高めるための一つの方針は，語順の考慮である [1, 4, 5]．また，別の方針として単語埋め込みの利用がある [6, 7]．CDER（Cover Disjoint Error Rate）[8] や EED（Extended Edit Distance）[9] などのジャンプ付き編集距離は語順の入れ替えを考慮する点で妥当だと考えられる．我々は語順の考慮と単語埋め込みの利用を両立するべく，ジャンプ付き編集距離の置換コストを単語埋め込みを用いて緩和した WCDER（Word embedding based CDER）を提案する．

問題設定　本論文中の自動評価指標は式 (1) のように単語数がそれぞれ n，n' の翻訳文（c）と参照文（r）を入力として，その類似度を計算して翻訳文の良し悪しを測定する．

$$c = [c_1, c_2, \cdots, c_n],\quad r = [r_1, r_2, \cdots, r_{n'}] \tag{1}$$

2　編集距離に基づく自動評価指標

この節では提案手法の基礎となる，編集距離（Edit Distance, ED）[10] とそれに基づく自動評価指標の先行研究を説明する．なお，本論文における編集距離は文字レベルではなく単語レベルの編集距離を意味する．編集距離は，2 つの単語列を入力とし，置換，削除，挿入の操作によって片方の単語列をもう一方の単語列に変換するのに必要な手順の最小コストとして定義され，通常各操作のコストは 1 である．編集距離は単語列ペアの非類似度を表す．つまり値が大きいほどペアの類似度は低いことを表す．編集距離の操作コストの総和を表す配列を $D^{|c|\times|r|}$ とすると，(2) 式で計算することができる．

$$D(i,j) = \min \begin{cases} D(i-1, j-1) + (1 - \delta(c_i, r_j)), \\ D(i-1, j) + 1, \\ D(i, j-1) + 1 \end{cases} \tag{2}$$

ただし，$D(1,1)$ は 0 であり，$\delta(w_i, w_j)$ は 2 単語 w_i，w_j が同じ場合は 1 を返し，それ以外の場合は 0 を返す．なお，最終的に得られる $D(|c|, |r|)$ が編集距離を表す．編集距離の計算と同時に，配列 $T^{|c|\times|r|}$ にどの操作により配列 D が更新されたのかを保存しておくことによって，$T(|c|, |r|)$ からバックワードすることでアラインメントを獲得できる．

2.1　ジャンプ付き編集距離: CDER

CDER は編集距離をジャンプ操作の追加によって拡張したもので，再帰式 (3) で表される．ただし $D(1,1)$ は 0 である．

$$D(i,j) = \min \begin{cases} D(i-1, j-1) + (1 - \delta(c_i, r_j)), \\ D(i-1, j) + 1, \\ D(i, j-1) + 1, \\ \min\limits_{i'} D(i', j) + 1 \end{cases} \tag{3}$$

ここで，$D(i,j)$ を計算するにあたって，各 j について以下の 3 ステップの計算が必要となる．

1. それぞれの i で，式 (3) の上 3 項の最小項を求める
2. $\min\limits_{i'} D(i', j)$ を求める

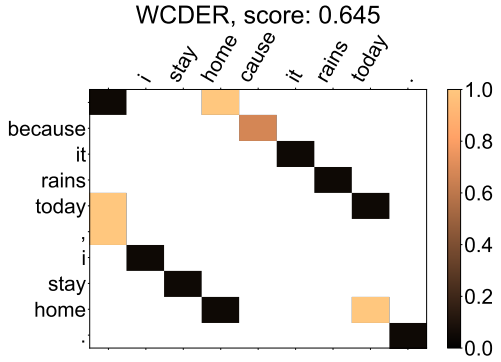

図 1 WCDER のアラインメント行列．縦軸，横軸はそれぞれ参照文と翻訳文．実行された編集操作のコストが高い場合は薄い色で，低い場合は濃い色で表されている．

3. それぞれの i で，式 (3) を求める

2.2 単語埋め込みに基づく編集距離: WED

WED（Word Embedding based Edit Distance）[11] は編集距離を，単語埋め込みを用いた操作コストの緩和によって拡張したものである．先行研究 [11] では全ての編集操作のコストを 1 から緩和しているが，本論文では置換コストのみを緩和しており，再帰式 (4) で表される．ただし $D(1,1)$ は 0 である．subcost は置換コストを表し，置換したい 2 単語の埋め込みのコサイン類似度が 0.5 以下の単語はほぼ類似性のない単語としてコスト 1，同単語はコスト 0 とし，その間を連続的にとるよう式 (5) で定義した．

$$D(i,j) = \min \begin{cases} D(i-1,j-1) + \mathrm{subcost}(c_i, r_j), \\ D(i-1,j) + 1, \\ D(i,j-1) + 1 \end{cases} \tag{4}$$

$$\mathrm{subcost}(c_i, r_j) = \frac{(1-0.5) - \max(0, \mathrm{sim}(c_i, r_j) - 0.5)}{1-0.5} \tag{5}$$

ここで，$\mathrm{sim}(w_i, w_j)$ は 2 つの単語 w_i,w_j の埋め込み $\boldsymbol{w_i}$,$\boldsymbol{w_j}$ のコサイン類似度を表し，式 (6) で定義される．

$$\mathrm{sim}(w_i, w_j) = \begin{cases} \cos(w_i, w_j) & (\boldsymbol{w_i} \& \boldsymbol{w_j} \text{ exist}) \\ 0 & (\text{else}) \end{cases} \tag{6}$$

3 提案手法: WCDER

本論文では，2.1，2.2 節で述べた拡張を組み合わせた WCDER を提案する．ジャンプ操作の追加は語順の入れ替えに対応し，単語埋め込みによる置換コストの緩和は文中の単語の類似度を考慮する．よって，これらの拡張を組み合わせることでより質の高い自動評価指標を作れると考えた．再帰式は

表 1 実験で用いた自動評価指標と各指標が考慮する項目

評価指標名	語順	単語類似度
BoW	×	✓
Vec Sum	×	✓
ED	✓	×
CDER	✓（ジャンプ付き）	×
WED	✓	✓
WCDER（提案手法）	✓（ジャンプ付き）	✓

式 (7) で表される．ただし，$D(1,1)$ は 0 で，2.1 節の CDER 同様各 j について 3 ステップの計算が必要になる．

$$D(i,j) = \min \begin{cases} D(i-1,j-1) + \mathrm{subcost}(c_i, r_j), \\ D(i-1,j) + 1, \\ D(i,j-1) + 1, \\ \min_{i'} D(i',j) + 1 \end{cases} \tag{7}$$

WCDER の実行の様子を図 1 に示した．縦軸，横軸に並んだ単語列はそれぞれ参照文と翻訳文を表す．置換は斜め，挿入は下，削除は右方向に対応づけられ，アラインメントが飛んでいる部分でジャンプが生じている．色が濃い部分は実行された操作のコストが低く，薄い部分はコストが高いことを表す．図 1 よりこの例文のアラインメントはジャンプ 3 回，置換 8 回，挿入 1 回でとられることが分かる．

4 実験

4.1 実験設定

語順と単語類似度の考慮が自動評価指標と人手評価の相関向上に寄与するかを確かめるため，表 1 に示した自動評価指標を用意して実験を行った．

データセット WMT19 Metrics Shared Task [1] のテストセット（人手評価で Better と評価された翻訳文，Worse と評価された翻訳文，参照文の 3 文が与えられる）を用いた．

評価尺度 Kendall の τ [2] の順位相関係数 [12] を用いた．

単語埋め込み 本論文全体で，単語埋め込みについては glove.840B.300d [3] を用いた．

ベースライン手法 Bag of Words のコサイン類似度（BoW）とベクトル和のコサイン類似度（Vec Sum）を用いた．

1) https://www.statmt.org/wmt19/metrics-task.html
2) Kendall の τ についての詳細は付録 A に示した．
3) https://nlp.stanford.edu/data/glove.840B.300d.zip

表 2　WMT19 Newstest19 における，DARR で測定したセグメントレベルの人間との順位相関．Kendall の τ の順位相関係数を使用している．言語対として，ドイツ語（de），フィンランド語（fi），グジャラート語（gu），カザフ語（kk），リトアニア語（lt），ロシア語（ru），中国語（zh）をそれぞれ英語（en）に翻訳したものを用いた．

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Average
#Sentences	85365	38307	31139	27094	21862	46172	31070	40182
BoW	0.017	0.218	0.184	0.365	0.230	0.107	0.297	0.203
Vec Sum	**0.090**	**0.252**	**0.213**	0.379	0.240	**0.145**	**0.340**	**0.237**
ED	-0.089	0.085	0.034	0.249	0.137	-0.013	0.200	0.086
CDER	0.021	0.220	0.167	0.367	0.246	0.104	0.311	0.205
WED	0.030	0.198	0.165	0.333	0.235	0.099	0.279	0.191
WCDER	0.065	0.243	0.207	**0.388**	**0.271**	0.133	0.332	0.234

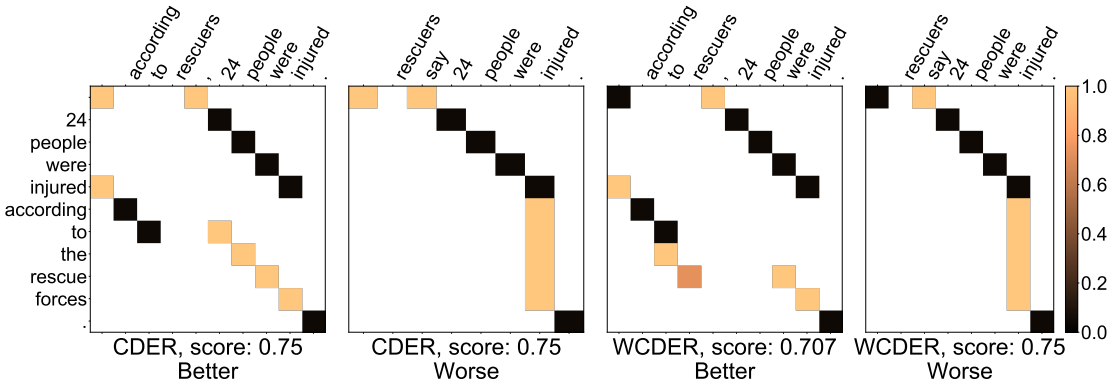

図 2　単語埋め込みによる置換コスト緩和が効果的だった例．縦軸が参照文で横軸が翻訳文．人手評価で良い翻訳文だったものが Better，悪かったものが Worse とラベル付けされている．

BoW は形式的に式 (8) のように定義される．

$$\mathrm{BoW}(c,r) = \cos\left(\sum_{i=1}^{|V|} t_i \boldsymbol{e}^{(i)}, \sum_{j=1}^{|V|} t'_j \boldsymbol{e}^{(j)}\right) \tag{8}$$

ここで，cos の引数は文ベクトルを表し，V は参照文と翻訳文に含まれる全単語の集合，t_i は V 中の i 番目の単語が c 中で現れた回数，t'_j は V 中の j 番目の単語が r 中で現れた回数，$\boldsymbol{e}^{(i)}$ は i 番目の値だけが 1 になっている長さ $|V|$ の one-hot ベクトルである．

Vec Sum は形式的に式 (9) のように定義される．

$$\mathrm{vec\ sum}(c,r) = \cos\left(\frac{\sum_{i=1}^{|c|} \boldsymbol{c_i}}{|c|}, \frac{\sum_{j=1}^{|r|} \boldsymbol{r_j}}{|r|}\right) \tag{9}$$

ここで，cos の引数は文ベクトルを表し，$\boldsymbol{c_i}$,$\boldsymbol{r_j}$ はそれぞれ c_i,r_j の単語埋め込みを表す．

前処理　前処理として，入力された文を単語ごとに分割し小文字化した．また，ジャンプ付き編集距離の先行研究 [8, 9] に基づき文頭に空白を追加した．

操作コスト　挿入，削除，ジャンプのコストは全て 1 とした．置換コストは単語埋め込みを用いない場合には 1 とし，用いる場合には式 (5) で求めた．

正規化　ジャンプがある指標については先行研究 [9] に記載のある式 (10) で，それ以外の指標では参照文の長さで割ることで正規化した．

$$\mathrm{Score} = \frac{(\text{sum of all costs}) + v}{|\text{reference}| + v} \tag{10}$$

ここで，v はペナルティ項で，翻訳文中の単語それぞれの | アライン回数 − 1| の和を表しており，アライン回数が 0 回，あるいは 2 回以上となる単語の数に応じてペナルティを課している．先行研究 [9] では v にハイパーパラメータ ρ がかけられているが，他の手法との差分をなくすため本論文では省いた．

4.2　結果

実験の結果を表 2 に示した．編集距離に基づく指標（ED, CDER, WED, WCDER）を見ると，ED と比べて，CDER および WED の相関が上昇した．また，CDER および WED と比べて，WCDER の相関が上昇した．よって．ジャンプ操作の追加と単語埋め込みの利用のいずれの拡張も人手評価との相関向上に寄与したことが確認できた．また，Vec Sum と比べて WCDER の大幅な相関向上は見られなかった．

4.3　議論

埋め込み利用の効果が出た例　de-en の 59504 文ペア目を CDER と WCDER で評価したものが図 2 で

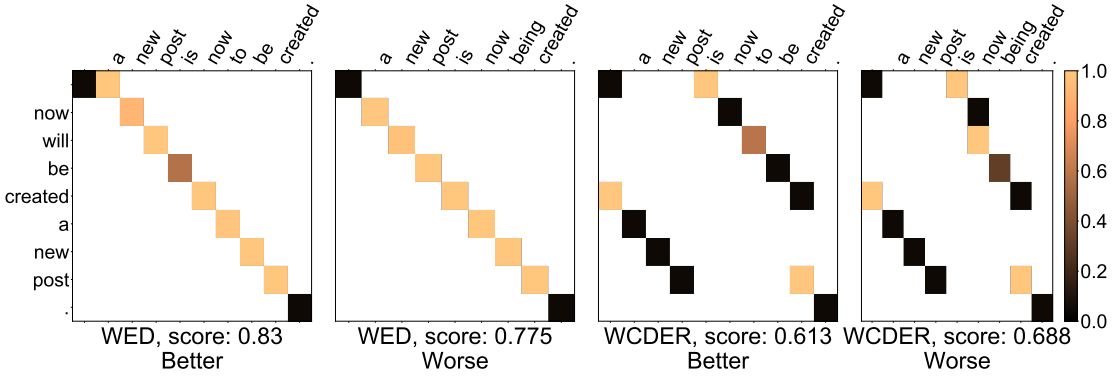

図 3　ジャンプ操作の追加が効果的だった文ペア例．縦軸が参照文で横軸が翻訳文．人手評価で良い翻訳文だったものが Better，悪かったものが Worse とラベル付けされている．

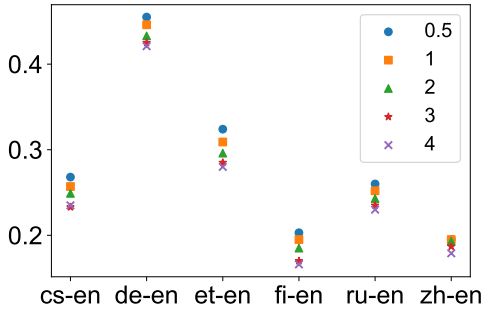

図 4　WCDER のジャンプコストを 0.5, 1, 2, 3, 4 として WMT18 Newstest18 を用いて 4 節同様の実験をした結果のプロット．横軸は言語対を，縦軸は順位相関を表す．

ある．CDER では 2 つの翻訳文のスコアが同じだった．しかし，WCDER では “rescuers” と “rescue” が対応づけられた結果 Better の文のスコアが良くなった．よって，単語埋め込みの利用により，単語の表層だけでなく意味を汲みとって評価したことが分かる．

ジャンプ操作追加の効果が出た例　同様に，de-en の 75584 文ペア目を WED と WCDER で評価したものが図 3 である．WED では Worse のスコアの方が良くなっていた．しかし，WCDER ではジャンプが生じて翻訳文の後半と参照文の前半で対応づけが起こった．それらの類似度が考慮された結果 Better の文のスコアが良くなった．よって，ジャンプにより語順の入れ替えに対応して評価したことが分かる．

ジャンプコスト　ジャンプコストは「意味が似た単語のかたまりがどれくらい大きければジャンプが生じるか」に対応する値だと考えられる．つまり，ジャンプコストが大きい場合，意味が似た単語のかたまりが大きいときに限りジャンプが生じる．反対にジャンプコストが小さい場合，意味が似た単語のかたまりが小さいときであってもジャンプが生じる．図 4 は WMT18 Newstest18 データセットを用いて，ジャンプコストを変更しながら 4 節同様の実験をした結果である．図 4 より，ジャンプコストが大きくなるにつれ WCDER の評価と人手評価の相関が下がっていることが分かる．相関低下の要因として，他の操作コストに比べて不当に大きなジャンプコストが語順が考慮を妨げている可能性が考えられる．したがって，人手評価と高い相関を持つ評価を行うためには，ジャンプコストとその他の操作コストとの間の適切なバランスが重要になると考えられる．また，4 節の実験で用いたベクトル和のコサイン類似度という指標は語順を考慮しない．この指標は，WCDER のジャンプコストを極端に小さくした指標とも捉えられる．表 2 より，ベクトル和のコサイン類似度と WCDER の結果はほとんど変わらなかった．そのため，削除や挿入のコストを 1 とした本論文の実験設定は，ジャンプコストとのバランスという観点で適切ではない可能性が示唆される．

今後の方向性　削除，挿入コストとジャンプコストとの適切なバランスを探りたいと考えている．さらに，そのバランスを保ちつつ削除，挿入コストも単語埋め込みを用いて緩和することで，提案手法をより良い指標へと発展させていきたい．

5　結論

本論文では編集距離をジャンプ操作と単語埋め込みによる置換コストの緩和で拡張した WCDER を提案した．WCDER は語順と単語類似度の考慮を両立した翻訳システムの自動評価指標である．この両立により人手評価との相関が上昇したことに加え，解釈性の高いアラインメント情報が得られる．これは機械翻訳システム開発に有用だと考えられる．提案手法の改善の余地は大きいと思われるため，今後も調査を継続していきたい．

謝辞

本研究は JSPS 科研費 JP22H05106，JST ACT-X JPMJAX200S，JST CREST JPMJCR20D2 の助成を受けたものです．また，本研究を進めるにあたり，頻繁に議論に参加していただいた東北大学乾・坂口・徳久研究室，東北大学鈴木研究室の皆様へ感謝いたします．

参考文献

[1] Tsutomu Hirao, Hideki Isozaki, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Evaluating translation quality with word order correlations. **Journal of Natural Language Processing**, Vol. 21, No. 3, pp. 421–444, 2014.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[3] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015.

[4] Hiroshi Echizen-ya and Kenji Araki. Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. In **Proceedings of Machine Translation Summit XI: Papers**, Copenhagen, Denmark, September 10-14 2007.

[5] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004.

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2019.

[7] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 563–578, November 2019.

[8] Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDER: Efficient MT evaluation using block movements. In **11th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 241–248, Trento, Italy, April 2006.

[9] Peter Stanchev, Weiyue Wang, and Hermann Ney. EED: Extended edit distance measure for machine translation. In **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 514–520, August 2019.

[10] Vladimir I Levenshtein, et al. Binary codes capable of correcting deletions, insertions, and reversals. In **Soviet physics doklady**, Vol. 10, pp. 707–710. Soviet Union, 1966.

[11] Yilin Niu, Chao Qiao, Hang Li, and Minlie Huang. Word Embedding based Edit Distance. **arXiv e-prints**, p. arXiv:1810.10752, October 2018.

[12] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 62–90, Florence, Italy, August 2019.

A　Kendall の τ の詳細

WMT19 Metrics Shared Task では，自動評価指標の評価を行うにあたって，Kendall の τ の順位相関を計算している．このタスクでは，「ある入力文に対して 2 つの機械翻訳システムが出力した翻訳文」，「2 つの翻訳文のうちどちらが Better か Worse か人手評価したラベル」，「参照文」が与えられる．自動評価指標によって 2 つの翻訳文と参照文の類似度をそれぞれ算出し，2 つの翻訳文のうちどちらの翻訳文が Better か Worse かを評価する．そして，2 つの翻訳文に対する人手評価と自動評価指標による評価が一致していれば Concordant（$Conc$），不一致であれば Discordant（$Disc$）となる．こうして全ての文ペアについて $Conc$ か $Disc$ かを求めた後，それらの個数を用いて式 (11) より τ を計算する．

$$\tau = \frac{|Conc| - |Disc|}{|Conc| + |Disc|} \tag{11}$$