文埋め込みモデルの内部表現と不均衡最適輸送を用いた機械翻訳自動評価

蒔苗茉那　五藤巧　坂井優介　上垣外英剛　渡辺太郎
奈良先端科学技術大学院大学
{mana.makinae.mh2, goto.takumi.gv7,
sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

機械翻訳をはじめとする自然言語生成タスクにおいて，自動評価は表層的な手法からニューラルベースの手法へと発展してきたが，評価値がなぜその結果に至ったのか解釈性に課題がある．提案手法では，文埋め込みモデルの内部表現と不均衡最適輸送をアライメントとして用いることで，参照文と生成文の間における部分的かつ非対称な対応関係を獲得する．これにより，柔軟なトークン間の対応関係の可視化が可能となり，評価値の解釈性を向上させる．実験結果から，従来の BERT に基づく表現を文埋め込みモデルの内部表現に置き換えるだけでも，人手評価との相関が改善される場合があることが確認されたことから，アライメント手法そのものよりも，文埋め込みモデルの内部表現を用いることの寄与が相対的に大きい可能性を示唆している．

1　はじめに

自動評価は，機械翻訳をはじめとする自然言語生成タスクにおいて不可欠な役割を果たしてきた [1, 2, 3, 4]．BLEU [1] や ROUGE [5] といった表層的な n-gram 一致に基づく指標が長く用いられてきた一方で，COMET [2] や BLEURT [3] など，文脈埋め込み表現を活用したニューラルベースの自動評価も主流となっている．また，大規模言語モデルを自動評価として用いる手法 [6, 7] も提案されている．

一方で，これら自動評価は，解釈性の低さという課題を抱えている．ある出力が高い，あるいは低い評価を受けた時，その原因を把握することが難しく，特に COMET [2] など文全体を単一のスコアで評価する手法は，どの単語が評価に影響を与えたのかが不明瞭であり，分析への応用が制限される．

解決策のひとつにトークンレベルの対応を明示的に扱う BERTScore [8] がある．BERTScore は，文脈埋め込み空間におけるトークン間の類似度を測ることで，生成文と参照文の間の対応関係を可視化できる点で，ニューラルベースの自動評価の中でも比較的高い解釈性を有する．一方で，BERTScore が用いる埋め込み表現は主に BERT を用いたものであり，近年提案されている文単位の意味表現に特化した文埋め込みモデルの知見は十分に活用されていない．文単位の意味表現で高い性能を示す文埋め込みモデル [9, 10, 11] は，文単位の表現を目的として学習されているが，内部表現には単語・サブワードレベルの意味情報が反映されていると考えられ，トークンレベルで活用できる可能性がある．

また、BERTScore におけるトークン対応は，最大コサイン類似度に基づく貪欲的な対応付けによって決定されるため，対応関係が過度に強制されるため，生成文と参照文の情報量が不均衡な場合や Addition や Omission といった翻訳特有の誤りが生じる場合に，それらを適切に扱うことが難しい．

そこで本研究では，文埋め込みモデルの内部表現と，不均衡最適輸送 [12, 13] を用いてトークン間のアライメントを推定する [14, 15, 16]．これにより，部分的あるいは非対称な対応関係を許容し，より柔軟で情報量の多いアライメントを獲得を目指す．実験では，WMT 2022 Metrics Shared Task [17] を元に提案手法を評価した．その結果，特定の言語対においては，既存の自動評価と比較して人手評価との相関が向上することを確認した．

2　関連研究

2.1　BERTScore

BERTScore [8] は，文脈埋め込みを用いて生成文と参照文の意味的類似度を評価する自動評価であ

る．従来の BLEU[1] や ROUGE [5] に代表される表層的な n-gram 一致に基づく指標とは異なり，事前学習済み言語モデルから得られるトークン埋め込みを用いることで，語彙的に異なるが意味的に近い表現を捉えることができる．具体的には，参照文 $\boldsymbol{x}$ および生成文 $\hat{\boldsymbol{x}}$ に含まれる各トークン間のコサイン類似度を計算し，各トークンが最も類似するトークンと対応付けられる貪欲的なマッチングに基づいて，Precision，Recall，および F_1 スコアを算出する．

$$R_{\text{BERT}} = \frac{1}{|\boldsymbol{x}|} \sum_{x_i \in \boldsymbol{x}} \max_{\hat{x}_j \in \hat{\boldsymbol{x}}} x_i^\top \hat{x}_j, \tag{1}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{\boldsymbol{x}}|} \sum_{\hat{x}_j \in \hat{\boldsymbol{x}}} \max_{x_i \in \boldsymbol{x}} x_i^\top \hat{x}_j, \tag{2}$$

$$F_{1,\text{BERT}} = \frac{2\, P_{\text{BERT}}\, R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{3}$$

さらに，IDF に基づく重み付けを導入することで，意味的に重要なトークンが評価値により大きく反映されるようになっている．以上から，BERTScore はトークンレベルの対応関係を明示的に扱う点で，文全体を単一のベクトルとして扱う評価指標よりも高い解釈性を有しているといえる．

2.2 最適輸送

最適輸送（Optimal Transport; OT）[18] は，過不足のないアライメントを通じて点群間の確率分布間の距離を測る手法である．自然言語処理においては，文中のトークンを分布として捉え，トークン間の意味的距離を輸送コストの問題に置き換えることで，意味的な距離の計算が可能となる．このとき，意味的に近い語同士は低コストで輸送でき，意味的に離れた語同士は輸送コストが高くなる．最適輸送問題は，次式で定義される．

$$\text{OT}(\mathbf{a}, \mathbf{b}, \mathbf{C}) = \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \sum_{i,j} P_{ij} C_{ij}, \tag{4}$$

$$\text{s.t. } U(\mathbf{a}, \mathbf{b}) = \left\{ \mathbf{P} \,\middle|\, \mathbf{P} \in \mathbb{R}_+^{n \times m},\ \mathbf{P}\mathbf{1} = \mathbf{a},\ \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}. \tag{5}$$

ここで，$\mathbf{a} = (a_1, \ldots, a_k)$ および $\mathbf{b} = (b_1, \ldots, b_m)$ は，それぞれ 2 つの点集合上の確率分布を表す．a_i および b_j は，それぞれ分布 $\mathbf{a}$ および $\mathbf{b}$ における i 番目，j 番目の確率分布である．$\mathbf{P}$ は輸送計画の候補を表し，P_{ij} は点 a_i から b_j への輸送量を示す．また，C_{ij} は a_i と b_j の間の輸送コストを表す．

2.3 不均衡最適輸送

OT は，点群間の確率分布が過不足がないことを前提としており，すべての点が必ず対応付けられるという制約を持つ．しかし，翻訳や要約などの自然言語生成タスクにおいては，この仮定は必ずしも適切ではない．実際には，対応する語を持たないトークンや，一対多・多対一の対応関係 [19] が頻繁に生じるからである．先行研究でも、BERTScore の枠組みで最適輸送を用いた対応付けを試みたものの，必ずしも性能向上には至らなかったことも報告されている [8]．不均衡最適輸送（Unbalanced Optimal Transport; UOT） [20] は，過不足の量に応じてペナルティを与えることで制約を緩和する．UOT は，次のように定式化される．

$$\begin{aligned}\text{UOT}(\mathbf{a}, \mathbf{b}, \mathbf{C}) = \min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \sum_{i,j} P_{ij} C_{ij} \\ + \lambda_1 \,\text{KL}(\mathbf{P}\mathbf{1}_m \,|\mathbf{a}) + \lambda_2 \,\text{KL}(\mathbf{P}^\top \mathbf{1}_m \,|\mathbf{b}).\end{aligned} \tag{6}$$

ここで，$\mathbf{P} \in \mathbb{R}_+^{n \times m}$ は，必ずしも与えられた周辺分布を厳密に満たす同時確立である必要はないことを示す．その代わりに，周辺分布からの乖離を Kullback–Leibler（KL）ダイバージェンスによる正則化項として導入し，分布間の不一致をペナルティとして扱う．このとき，λ_1 および λ_2 は，輸送量の不一致に対するペナルティの強さを制御するハイパーパラメータである．

3 提案手法

本研究では，BERTScore に着想を得て，評価性能と解釈性の向上を目的とした新たな自動評価を提案する．まず，文埋め込みモデル [9, 10, 11] から得られる内部表現をトークンとして利用する．その上で，不均衡最適輸送を用いて，生成文と参照文のトークン間のアライメントを推定する．これにより，すべてのトークンを強制的に対応付けることなく，生成文と参照文の間に生じる非対称な対応関係や文長の違いを柔軟に扱い、それらをスコア計算に反映することで性能向上を図る．

文埋め込みモデル の内部表現表現　文埋め込みモデル は，文全体の意味表現を獲得することを目的として学習されたモデル [9, 10, 11] であり，通常は文単位での意味的対応付けに用いられる．一方で，文の意味表現を構築する過程において，各単語の意

味的差異や文中での役割を内部的に捉えていると考えられる．すなわち，文単位の意味表現を学習する過程で，単語レベルの情報も同時に反映された表現が獲得されている可能性がある．本研究では，このような 文埋め込みモデル の性質に着目し，文全体を単一のベクトルに集約する前の中間表現を用いることで，内部表現単位の表現を取得し，それらをトークン間アライメントの推定に利用する．

UOT によるトークンアライメント　参照文および生成文を，文埋め込みモデルの内部表現を用いてトークンに変換する．参照文のトークン埋め込み列を $X = \{x_1, \ldots, x_k\}$，生成文のトークン埋め込み列を $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_m\}$ と表す．ここで，$x_i, \hat{x}_j \in \mathbb{R}^d$ は，それぞれ参照文および生成文に含まれるトークンの d 次元文脈埋め込みを表す．

UOT に基づく評価スコア　トークン間アライメントを推定するために，参照文および生成文のトークン埋め込みから輸送コスト行列 $\mathbf{C}$ を構成し，不均衡最適輸送（UOT）に基づいて輸送計画 $\mathbf{G}$ を求める．各トークンの重み $\mathbf{a}, \mathbf{b}$ は，トークン埋め込みの L_2 ノルムに基づいて次のように定義する：

$$a_i = \|x_i\|_2, \quad b_j = \|\hat{x}_j\|_2. \tag{7}$$

これは，情報量の大きいトークンにより大きな重みを与えることを意図している．輸送計画 $\mathbf{G}$ は，不均衡最適輸送問題を解くことで $\mathbf{G} = \mathrm{UOT}(\mathbf{a}, \mathbf{b}, \mathbf{C})$ として求める．輸送コスト行列 $\mathbf{C}$ は，生成文および参照文のトークン埋め込み間のユークリッド距離として定義する．得られた輸送計画に基づき，五藤ら [21] に従い、輸送量の総和を真陽性（TP）として解釈し，トークン重みとの差分から偽陽性（FP）および偽陰性（FN）を次のように定義する：

$$\mathrm{TP} = \sum_{i=1}^{k} \sum_{j=1}^{m} G_{ij}, \tag{8}$$

$$\mathrm{FP} = \sum_{i=1}^{k} a_i - \mathrm{TP}, \tag{9}$$

$$\mathrm{FN} = \sum_{j=1}^{m} b_j - \mathrm{TP}. \tag{10}$$

これらを用いて，BERTScore と同様の枠組みで Precision，Recall，および F_1 スコアを算出する：

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{11}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{12}$$

$$F_1 = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{13}$$

4　実験設定

提案手法の有効性を検証するため，WMT 2022 Metrics Shared Task のデータセットを用いて評価を行った．本タスクでは，人手評価に基づく複数の評価値が提供されており，自動評価の性能を比較するためのベンチマークとして用いられている．

人手評価　評価には, MQM（Multidimensional Quality Metrics）および DA（Direct Assessment）の二種類の人手評価を用いた．MQM については en–de, en–ru の言語対を対象とし，DA については cs–en, ja–en，ru–en，de–en，uk–en の言語対を対象とした．

評価指標　自動評価の性能は，WMT 2022 Metrics Shared Task の設定に従い，人手評価との Pearson 相関係数によりシステムレベルで評価した．また分析の際にはセグメントレベルでも評価している．

比較手法　ベースラインとして，BERTScore で用いられているトークン間のコサイン類似度に基づく最大類似度選択による貪欲マッチングで得たアライメントを元にスコアを算出する．提案手法では，この貪欲マッチングの代わりに，UOT よりトークン間の対応関係を推定し，得られた輸送計画に基づいてスコアを算出する．

文埋め込みモデル　本研究では，以下の文埋め込みモデルの内部表現を用いた：(i) Language-agnostic BERT 文埋め込みモデル (LaBSE) [9]，(ii)Multilingual E5 Text Embeddings (E5) [10]，(iii) Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation (M3) [11]．ベースラインおよび提案手法では，いずれも同一の文埋め込みモデルを用いることで，アライメント手法の違いの影響を比較した．また，実験では，文埋め込みモデルの内部表現のうち、最終層から得られる表現を用いている．

UOT のハイパーパラメータ　UOT の正則化パラメータとして λ_1 および λ_2 を用い，それぞれ 0.1 から 1.0 の範囲で 0.1 ずつ変化させながら探索し，オラクル性能を報告する．

5　実験結果

図 1 は、en–de および en–ru において、同一の文埋め込みモデルを用いた条件下で，既存の BERTScore における貪欲的アライメントと，提案手法における UOT に基づくアライメントによる効果を比較した結果である．他の言語ペアの結果は付録 A に示す。

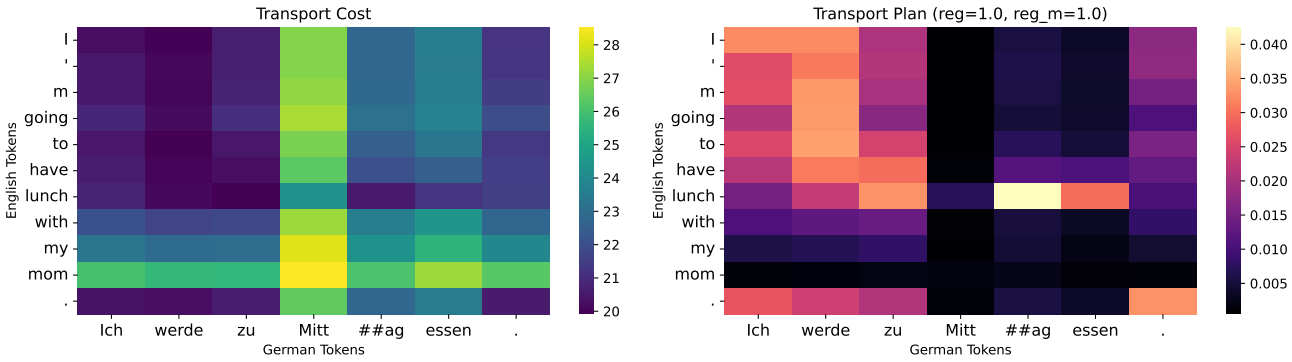

図 1　英独のトークン間アライメントのヒートマップ

表 1　WMT 2022 Metrics Shared Task における人手評価とのピアソン相関．HT は，評価セットに人間による翻訳が含まれているかどうかを，() は用いたモデルを示す．

Metric	en–de		en–ru
	w/ HT	w/o HT	w/o HT
COMET-22	0.761	0.771	0.900
COMET-20	0.812	0.876	0.936
Greedy (bert)	0.338	0.428	0.811
Greedy (LaBSE)	–	0.608	0.746
Greedy (E5)	–	0.611	0.783
Greedy (M3)	–	0.648	0.815
UOT (LaBSE)	–	0.425	0.852
UOT (E5)	–	0.455	0.862
UOT (M3)	–	0.438	0.876

結果から，en–ru は，提案手法が，貪欲的アライメントによる測り方よりも人手評価との高い相関を示した．一方で，en–de は，既存の BERTScore で用いられるアライメント手法の方が，提案手法よりも高い相関を示した．これらの結果から，従来の BERT から文埋め込みモデルの内部表現に置き換えるだけでも，人手評価との相関が改善される場合があることが確認された．また，提案手法は，言語対によっては性能向上を示す一方で，必ずしも一貫して優位となるわけではないことも明らかとなった．これは，提案手法における性能向上の要因として，アライメント手法そのものよりも，文埋め込みモデルをトークン表現として用いることの寄与が大きい可能性を示唆している．

6　分析と考察

トークン間アライメントを可視化したヒートマップを分析すると，en–de においても提案手法が意味的に妥当なトークン対応を捉えていることが確認できる．図 1 に示すように，英語文 “I’m going to have lunch with my mom” に対して，ドイツ語文 “Ich werde zu Mittag essen”（“I’m going to have lunch” に対応）を与えた場合，“with my mom” に対応する英語側トークンには，輸送量がほとんど割り当てられていない．これは，対応先を持たない情報が適切に抑制されていることを示しており，UOT に基づくアライメントが機能している可能性を示唆している．そのため，en–de における人手評価との相関低下は，アライメント自体の品質に起因するものではない可能性がある．なお，en–ru の場合を付録 B に示す．

7　おわりに

本研究では，文埋め込みモデルの内部表現と UOT に基づくアライメントを使用することで，参照文と生成文の間のトークンレベルの対応関係を柔軟かつ明示的に捉えることが可能な，学習を必要としない多言語対応の自動評価を提案した．実験結果から，一部の言語対においては，提案手法が既存の BERTScore に基づく手法と比較して，人手評価とより高い相関を確認した．同時に，アライメント手法そのものよりも，文埋め込みモデルの内部表現を用いることの寄与が相対的に大きい可能性が示唆された．また，ヒートマップによるアライメントの分析結果から，トークン間アライメントが意味的に妥当である場合であっても，人手評価との相関で既存手法を下回ることも明らかとなった．今後の課題として，本研究で用いた文埋め込みモデルについても，文全体の意味表現を目的として学習されたモデルでありながら，単語レベルの意味的差異をどの程度正確に捉えているのかを分析するとともに，提案手法を最大限に活かすために，文埋め込みモデルがどのような性質を満たすべきかについてさらなる検討を行う．さらに，要約や画像キャプション生成の評価にも適用し，手法の有効性を検証していく．

謝辞

本研究の一部は JST SPRING JPMJSP2140 の助成を受けたものである.

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 Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[3] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[4] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

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

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[7] Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 229–246, 2024.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[9] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[11] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. NIPS'15, p. 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

[13] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. **Mathematics of Computation**, Vol. 87, , 07 2016.

[14] Kyle Swanson, Lili Yu, and Tao Lei. Rationalizing text matching: Learning sparse alignments via optimal transport. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5609–5626, Online, July 2020. Association for Computational Linguistics.

[15] Weijie Yu, Liang Pang, Jun Xu, Bing Su, Zhenhua Dong, and Ji-Rong Wen. Optimal partial transport based sentence selection for long-form document matching. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 2363–2373, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[16] Zihao Wang, Datong Zhou, Ming Yang, Yong Zhang, Chenglong Rao, and Hao Wu. Robust document distance with wasserstein-fisher-rao metric. In Sinno Jialin Pan and Masashi Sugiyama, editors, **Proceedings of The 12th Asian Conference on Machine Learning**, Vol. 129 of **Proceedings of Machine Learning Research**, pp. 721–736. PMLR, 18–20 Nov 2020.

[17] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[18] Leonid V Kantorovich. On the translocation of masses. **Manage. Sci.**, Vol. 5, No. 1, p. 1–4, October 1958.

[19] Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced optimal transport for unbalanced word alignment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3966–3986, Toronto, Canada, July 2023. Association for Computational Linguistics.

[20] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[21] 五藤巧, 坂井優介, 渡辺太郎. 文法誤り訂正における編集ベクトルの最適輸送に基づく性能評価尺度. 言語処理学会 第32回年次大会, 2026.

A　その他言語ペアでの実験結果

表 2　WMT 2022 における人手評価（DA）とのシステムレベルの Pearson 相関係数。テストセットには人手翻訳は含まれていない。

Metric	cs–en	de–en	ja–en	ru–en	uk–en
COMET-22	.821	.446	.976	.857	.714
COMET-20	.827	.424	.989	.847	.723
BERTScore (bert)	.825	.440	.988	.851	.717
BERTScore (LaBSE)	.816	.427	.993	.839	.701
BERTScore (E5)	.845	.413	.993	.783	.673
BERTScore (M3)	.840	.438	.993	.841	.693
Proposed (LaBSE)	.806	.483	.988	.824	.748
Proposed (E5)	.829	.523	.974	.831	.749
Proposed (M3)	.812	.582	.948	.803	.743

B　英露におけるヒートマップを用いたアライメントの分析

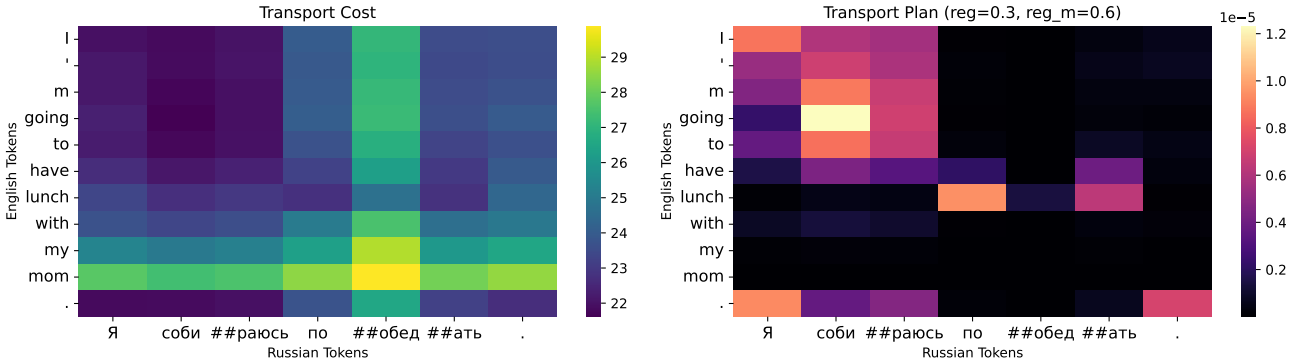

図 2　英露におけるトークン間アライメントのヒートマップ

図 2 は，英露（en–ru）におけるトークンアライメントのヒートマップを示す．英語文 “I’m going to have lunch with my mom” に対し，ロシア語文 “Я собираюсь пообедать.” を対応させた場合，“with my mom” に対応する英語側トークンに対しては，輸送量がほとんど割り当てられていないことが確認できる．これは，対応先を持たない情報が適切に無視されていることを示しており，不均衡最適輸送に基づくアライメントが意図どおり機能している可能性を示唆している．このように，en–de の場合と同様に，en–ru においてもトークン間アライメント自体は意味的に妥当な対応関係を捉えている．また en–ru では，評価スコアとしても人手評価との相関が高く，アライメント結果が評価指標として有効に機能していることが考えられる．