自然言語生成タスクの自動評価指標のためのドメイン外検出

伊藤拓海[1]　森下睦[2]　鈴木潤[1]
[1]東北大学　[2]NTT コミュニケーション科学基礎研究所
t-ito@tohoku.ac.jp　makoto.morishita.gr@hco.ntt.co.jp　jun.suzuki@tohoku.ac.jp

概要

自然言語生成タスクの評価指標として，ニューラルベースの評価指標が急速に発展している．しかし，これらの評価指標は，学習を伴う指標であり，評価するデータが評価指標の学習データのドメイン外の場合，誤ったスコアを付与する可能性がある．本稿では，機械翻訳のための参照なし評価指標である COMET-QE に対して，1 クラス分類器を評価指標に適用することで，ドメイン外検出をする方法を提案する．さらに，1 クラス分類器の分類精度向上を目指し，特徴量空間においてドメイン内とドメイン外のデータ点を分離できるように，評価指標の学習時に距離学習を適用した．実験の結果，既存手法であるモンテカルロドロップアウトを用いた手法を上回るドメイン外検出能力を達成した．

1　はじめに

機械翻訳や要約生成などの自然言語生成（NLG）タスクにおいて，NLG モデルの評価方法の研究が注目されている．その評価方法は，人手評価と自動評価に大きく分類される．人手評価は，人がモデルの性能を評価する方法であり，多くのタスクにおいてゴールドスタンダードとされている．しかしながら，人手評価は一般に時間的・金銭的なコストが高く，再現性の観点でも課題がある．そのため，代替手法として時間的・金銭的なコストが低く再現性も高い自動評価指標が用いられる．さらに，自動評価指標は参照あり評価指標と参照なし評価指標に大きく分類することができる．「参照」とは，各 NLG タスクの評価事例に対して人が作成した解答例となるテキストである．参照あり手法とは，参照テキストとシステムが生成したテキストを比べることで評価をする手法である．一方，参照なし評価指標とは参照テキストを使用しない評価方法である．つまり，参照なし評価指標は，評価事例として NLG モデルへの入力データとモデルの出力があれば適用可能である．新たなドメインの評価事例などに対して，より低コストで NLG モデルを評価できる可能性を秘めており，評価のスケールアップが期待されている．以前は，参照あり手法に比べて，評価性能が劣っているとされていたが，近年参照あり手法と競争的な評価性能を達成する手法がいくつか提案されてきている [1]. [1)] また，参照なし評価指標は NLG モデルの評価だけでなく，NLG モデルの推論時のリランキングなどにも使用されるなど，その応用範囲も広い [2].

こうした参照なし評価指標の発展の背景には，ニューラルベースの教師あり評価指標の発展がある．[2)] 特に，人が付与した品質評価スコアを教師に，BERT などの言語モデルをベースとした回帰モデルを学習する手法がよく提案されている [3]．しかし，こうした学習が伴う手法では，機械学習一般の課題であるが，学習データ中に評価事例の類似事例がない場合に，その評価事例の評価性能（或いは評価の信頼性）が低くなるという課題がある．実際，機械翻訳の多くの評価指標が，ドメインに対して頑健でないことが報告されている [4]．さらに，学習済みのモデルだけが配布されることが多々あり，モデルの学習に使用したデータの詳細が評価指標の利用者からは特定できない場合も少なくない．評価指標の使用方法を誤ったり，本来使用すべきではないデータに対して使用すると，NLG モデルの開発の失敗に繋がることや，誤った研究の結論を導く可能性がある．

本稿では，評価指標に対して評価するデータが学習データに含まれるドメインかどうかを検出する（ドメイン外検出と呼ぶ）機能を組み込むことを目指す．参照なし評価指標は学習データ外のドメインのデータに対して適用するされる可能性が高いため，本稿では参照なし評価指標に焦点を当てる．評

1）　NLG の評価指標は一般に，人手評価と相関（人手相関）が高いものほど高性能な指標とされている．
2）　ニューラルベースの NLG 評価指標は参照あり評価指標においても主流となっている．

価指標としての性能を維持したままドメイン外検出を実現することができれば，より公平で正確な NLG モデルの評価につながる．さらに，ドメイン外検出は評価指標の解釈性 [5] の向上にも貢献しうる．

2 関連研究

COMET [3] とは，機械翻訳タスクのための評価指標のフレームワークであり[3]，参照ありと参照なしの両方のバージョンを提供している（参照なしを COMET-QE [1] と呼ぶ）．ここでは，COMET-QE [1] について説明する．参照なし手法は，原文 s と，NLG システムの出力文 h を COMET-QE の入力とする．COMET-QE のモデル構造は，まず，XLM-RoBERTa [6] などの多言語マスク言語モデルによって原文と NLG システムの出力文から特徴量ベクトル $\boldsymbol{x}$ を得て，それをフィードフォワード層 $f(\cdot)$ に入力し，評価スコアを出力する．特徴量ベクトル $\boldsymbol{x}$ は以下のように計算される．

$$\begin{aligned} \mathbf{s} &= \text{XLM-RoBERTa}(s), \\ \mathbf{h} &= \text{XLM-RoBERTa}(h), \\ \mathbf{x} &= [\mathbf{h}; \mathbf{s}; \mathbf{h} \odot \mathbf{s}; \mathbf{h} - \mathbf{s}] \end{aligned} \tag{1}$$

人が付与するスコアを q，バッチサイズを n とすると，損失関数 L は以下のように計算される．

$$L = \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - q_i)^2 \tag{2}$$

本研究と類似し，NLG 評価指標に対して，信頼性スコアを出力する取り組みが既になされている [7, 8]．ドメイン以外にも，学習データや参照テキストのノイズなども不確定性の要因となり得る．例えば，Glushkova ら [7] は，モンテカルロドロップアウト (MCD) とディープアンサンブルの 2 つの不確実性推定手法を，COMET に適用している．MCD とは，推論時にもドロップアウトを適用し，複数回推論を行い，スコア集合 $Q = \{q_1, q_2, ..., q_i\}$ を求める．ガウス分布を仮定し，スコア集合 Q から，平均と標準偏差を推定する．この平均を新たなスコアとし，標準偏差を信頼性スコアとする．ディープアンサンブルは，複数のモデルを用意し，スコア集合を得る方法である．MCD やディープアンサンブルでは，推論に時間を要するという課題がある．Zerva ら [8] は，平均と標準偏差を直接出力するように COMET を学習する手法を提案している．なお，これらの論文は，主に参照ありの設定で実験を行なっている．

3) https://github.com/Unbabel/COMET

3 提案手法：学習ドメイン外検出

本研究の目標は，評価指標としての性能を維持したまま，ドメイン外検出を行うことである．本研究では，評価指標の学習データを正クラス，学習データ以外のドメインのデータを負クラスと呼ぶ．

3.1 1 クラス分類

ドメイン外検出のため 1 クラス分類器を適用する．1 クラス分類タスクとは正クラスかどうかを分類するタスクであるが，多クラス分類タスクとは異なり，学習時には正クラスのみのデータを使用し，推論時に正クラスか負クラスかを分類する．本実験では，正クラスのインスタンスの COMET-QE の特徴量ベクトルを 1 クラス分類タスクの学習データとする．つまり，COMET-QE の学習データを $D_{in-domain}$ とすると，1 クラス分類器の学習データ $D_{OneClass}$ は $D_{OneClass} = \{\mathbf{x}_i \mid (s_i, h_i) \in D_{in-domain}\}$ となる．また，この手法は評価指標の学習後に，1 クラス分類器の学習が可能であり，ドメイン外検出器を搭載することは，評価指標の性能には影響を与えない．なお，本稿では，1 クラス分類のアルゴリズムとして，高次元に対応した Empirical-Cumulative-distribution-based Outlier Detection (ECOD) [9] を使用する．

3.2 Triplet Margin Loss

本実験では，既存研究 [10] を参考に，1 クラス分類の精度向上を目指し，COMET-QE の学習時に距離学習を組み込み，特徴量埋め込みの空間において正クラスのデータ点を近づけ，正クラスと負クラスの距離を大きくする．本研究では，Triplet Margin Loss を使用する．なお，今回は負クラスのデータ点同士が近付くようには学習する必要がないため，アンカーには正クラスのサンプルだけを使用する．また，ミニバッチ内で作成できる全ての組み合わせを考慮する．つまり，アンカーを a，正クラスのインスタンスを p，負クラスのインスタンスを n，ミニバッチ内での組み合わせの数を c とすると，Triplet Margin Loss は以下のように計算される．

$$L_{triplet} = \frac{1}{c} \sum_{i}^{c} \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\} \tag{3}$$

なお，d はインスタンスの特徴量ベクトルの間の L2 距離である．

4 実験

第 3 節の手法の有効性を検証する．評価指標としての性能を落とすことなく，ドメイン外検出することを目指す．

4.1 モデル設定

本実験では，以下の 4 つの設定を比較する．

- MCD: COMET-QE に対して，MCD を適用 [7].
- OneClass: COMET-QE を学習し，その後 1 クラス分類器を学習．人手相関の観点では，COMET-QE に一致する．
- Margin(.05): COMET-QE の学習時に Triplet Margin Loss を margin 0.05 で学習し，その後 1 クラス分類器を学習．
- Margin(.1): COMET-QE の学習時に Triplet Margin Loss を margin 0.1 で学習し，その後 1 クラス分類器を学習．

本研究では，MCD の推論回数は 100 回に設定し，ドロップアウト率は 0.1 とした．また，本実験では，式 2 と Triplet Margin Loss を重み付けることなく線形に組み合わせた．ハイパーパラメータ等の詳細な設定は付録 A に示す．なお，第 2 節で紹介した，ディープアンサンブルや Zerva ら [8] の手法も比較可能な手法であるが，今後の課題とする．

4.2 学習/評価データセットと評価方法

学習データ． WMT 2022 Metrics Shared Task (WMT22) の学習データとして配布されている newstest2020 と newstest2021 の MQM データを使用する．[4] この MQM データには，英語-ドイツ語，中国語-英語，英語-ロシア語の 3 つの言語対が含まれている．なお，言語対によってスコアを付与した組織が異なり，スコアのレンジが異なる．[5] 本実験では，前処理として MQM スコアを正規化してから学習に使用した．また，Triplet Margin Loss で使用する負クラスのデータは ParaCrawl [11] から，言語対ごとに正クラスと負クラスのサンプルサイズが同数になるようにサンプリングした．MQM データの 1 割を開発データに使用し，残りの 9 割を COMET-QE の学習に使用した．

評価データ． WMT21 の TedTalks のデータセット [4] と WMT22 の評価用データセット [12] を使用する．どちらのデータも MQM でアノテーションされたデータであり，英語-ドイツ語，中国語-英語，英語-ロシア語の 3 つの言語対を含む．なお，WMT22 には news, conversational, e-commerce, social の 4 つのドメインが含まれている．本実験では，WMT22 の news を正クラスのドメイン，TedTalks と conversational, e-commerce, social の 4 つを負クラスのドメインとする．

評価方法． 評価指標の性能評価は，MQM スコアとの相関を報告する．[6] ドメイン外検出の評価には AUROC を使用する．

表 1 英語-ドイツ語の人手相関．値はセグメントレベルのケンドールの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec	social	tedtalks
MCD	**.236**	**.338**	.122	.294	**.266**	**.228**
OneClass	.233	.330	**.134**	.283	.251	.226
Margin(.05)	.232	.331	.132	**.297**	.224	.220
Margin(.1)	.146	.326	.100	.223	.191	.157

表 2 英語-ロシア語の人手相関．値はセグメントレベルのケンドールの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec	social	tedtalks
MCD	**.193**	**.309**	.159	.291	**.237**	**.168**
OneClass	.185	.301	.156	**.292**	.209	.165
Margin(.05)	.161	.288	**.172**	.264	.159	.152
Margin(.1)	.031	.257	.098	.218	.105	.092

表 3 中国語-英語の人手相関．値はセグメントレベルのケンドールの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec	social	tedtalks
MCD	**.346**	**.371**	**.257**	**.341**	**.322**	**.228**
OneClass	.343	.367	.245	.335	.321	.218
Margin(.05)	.333	.365	.228	.327	.308	.185
Margin(.1)	.318	.345	.174	.296	.309	.209

4.3 実験結果

表 1, 2, 3 に各データセットの人手相関（ケンドールの順位相関係数[7]）の結果を示す．“mixed” とは，

4) WMT22 では，Direct Assessments (DA) data という WMT の News translation task での人手評価のスコアも評価指標の学習データとして配布されている．先行研究 [1] などでは，DA data で事前学習したのち，MQM data で微調整をするという戦略をとっているが，本実験では DA data は使用しない．

5) 英語-ドイツ語と中国語-英語はスコアのレンジは $[-25, 0]$ である．一方で，英語-ロシア語はスコアのレンジは $[-\inf(-400), 100]$ である．

6) 相関の計算には https://github.com/google-research/mt-metrics-eval を使用した．相関の計算時には人手翻訳は含めていない．

7) ケンドールの順位相関係数は，セグメントレベルの MQM データセットの評価の際に使用される指標である．

表 4 AUROC．正クラスのデータは wmt22-news．

モデル	英語-ドイツ語				英語-ロシア語				中国語-英語			
	conv.	ec.	social	tedtalks	conv.	ec.	social	tedtalks	conv.	ec.	social	tedtalks
MCD	.559	.466	.552	.528	.686	.549	.635	.534	.233	.566	.463	.193
OneClass	.846	.779	.721	.722	.823	.753	.699	.692	.899	.756	.625	.639
Margin(.05)	.873	.798	.715	.722	**.865**	.785	.701	.705	.923	.773	.628	.674
Margin(.1)	**.910**	**.889**	**.817**	**.815**	.705	**.857**	**.789**	**.794**	**.946**	**.877**	**.767**	**.905**

wmt22 の 4 つのドメインを全て混合させた設定である．なお，表の conv. と ec はそれぞれ，conversational と e-commerce の略記である．表 4 には AUROC の結果を示す．付録 B にシステムレベルの評価結果とケンドールの順位相関係数以外の相関の結果を示す．なお，OneClass の人手相関は COMET-QE と一致する．

人手相関に関して，news と news 以外のドメインを比較すると，news 以外のドメインでは相関が弱くなっていることがわかる．例えば，表 1 の英語-ドイツ語では，COMET-QE（OneClass）は人手相関が wmt22-news では 0.330 であるのに対して，wmt22-conversation では，0.134 である．今回の実験結果からは，news 以外のドメインのテストデータがドメイン以外の問題で難易度が高くなっている可能性も捨てきれない．しかし，少なくとも，ドメインによっては相関が低い場合があることがわかる．

表 4 より，MCD ではほとんどの言語対・データセットで AUROC が 0.5 前後であり，ドメイン外検出ができていないことがわかる．いくつかの設定では，0.5 を下回っている（表の赤字）．特に，MCD は中国語-英語の wmt22-conversation，wmt22-social に対し，AUROC が 0.233，0.193 と低い値になっている．つまり，これらのデータセットでは，MCD では正のクラス wmt22-news の方が標準偏差の値が大きくなっていることを意味する．

一方，OneClass は一貫して，MCD を超えるドメイン外検出性能を達成しており，COMET-QE の特徴量ベクトルにドメインの情報が含まれていることが示唆される．特に，どの言語対でも conversation のドメインに対しては AUROC の値が 0.8 を超えている．conversation は特に人手相関が弱いドメインであり，1 クラス分類を適用することで，評価指標の評価性能が悪いドメインを検出できていることが示唆される．

また，Triplet Margin Loss を適用すると，一貫して AUROC の値が向上していることがわかる．しかし，ドメイン外検出の精度は向上したが，ほとんどのデータセットで人手相関が弱くなっている．特に，OneClass と比較すると，Margin が 0.1 の時に大きく相関が下がっている．また，正クラスである wmt22-news においても，一貫して相関が弱くなっている．距離学習を組み込むことで，特徴量空間において正クラスと負クラスのデータ点を分けることはできたものの，評価指標の評価性能自体に悪影響を与えてしまっていることが示唆される．別の距離学習アルゴリズムを使用したり，距離学習時に使用する負クラスのサンプルの選択方法を工夫したり，学習戦略やハイパーパラメータの微調整したりすることによって改善する可能性があると考えている．それらの検証は今後の課題である．

5 おわりに

本研究はドメイン外検出の機能を NLG 評価指標に適用することを目指したものである．本実験では，参照なし評価指標に焦点をあて，1 クラス分類器を参照なし評価指標に組み込むことを検証した．また，距離学習である Triplet Margin Loss を用いて，評価指標の特徴量空間において，学習ドメインとそれ以外のドメインのデータ点の距離を離すことを試みた．その結果，先行研究である MCD と比較して，1 クラス分類器を用いた手法が評価指標としての性能に影響を与えることなく，ドメイン外検出ができていることがわかった．また，Triplet Margin Loss を用いることで，ドメイン外検出の精度は向上したが，評価性能が低下してしまった．特に，正クラスのドメインに対しても人手相関が弱くなってしまっている．

本稿では，参照なし評価指標を対象としたが，本稿の手法は参照あり評価指標にも適用可能であり，今後の課題である．ドメイン外検出に焦点を当てたが，学習ドメイン外でも高性能な評価指標の実現のため，分布外一般化 [13] も，今後の研究の重要な方向性の一つである．NLG 分野において，高性能で頑健な自動評価指標の開発は重要な研究課題であり，今後更なる研究の推進を期待する．

謝辞

本研究は JSPS 科研費 JP21J14152, JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである.

参考文献

[1] Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 1030–1040, Online, November 2021. Association for Computational Linguistics.

[2] Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-Aware Decoding for Neural Machine Translation. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[4] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 733–774, Online, November 2021. Association for Computational Linguistics.

[5] Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. Towards Explainable Evaluation Metrics for Natural Language Generation, 2022.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[7] Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. Uncertainty-Aware Machine Translation Evaluation. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 3920–3938, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[8] Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. Disentangling Uncertainty in Machine Translation Evaluation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. arXiv, 2022.

[9] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. **IEEE Transactions on Knowledge and Data Engineering**, pp. 1–1, 2022.

[10] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive Out-of-Distribution Detection for Pretrained Transformers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 1100–1111, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics.

[12] Tom Kocmi, Rachel Bawden, OndÅ™ej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal NovÃ¡k, Martin Popel, MajaPopoviÄ‡, Mariya Shmatova. Findings of the 2022 Conference on Machine Translation (WMT22). In **Proceedings of the Seventh Conference on Machine Translation**, pp. 1–45, Abu Dhabi, December 2022. Association for Computational Linguistics.

[13] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey, 2021.

A 実験設定

表 5 COMET-QE のハイパーパラメータ

	設定
Encoder Model	XLM-RoBERTa (large)
Optimizer	AdamW
Learning rate	3.1e-05
Encoder Learning Rate	1.0e-05
Layerwise Decay	0.95
Batch size	16
Dropout (Feedforward)	0.15
Dropout (Encoder)	0.1
Hidden sizes	[2048, 1024]
Max Epochs	15

表 5 に COMET-QE の学習時のパラメータを示す．本実験では，Triplet Margin Loss を使用した際も同様の設定を用いた．毎エポック後にモデルを保存し，開発データに対するケンドールの順位相関係数が最も高かったチェックポイントのモデルを評価に使用した．なお，ケンドールの順位相関係数に対して，Early Stopping を行った．

B 実験結果

表 6 英語-ドイツ語のセグメントレベルのピアソンの相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.358	.537	.198	.443	.359	.324
OneClass	.375	.517	.258	.448	.363	.313
Margin(.05)	.384	.507	.256	.444	.337	.306
Margin(.1)	.297	.534	.177	.394	.266	.263

表 7 英語-ロシア語のセグメントレベルのピアソンの相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.272	.432	.240	.483	.355	.178
OneClass	.253	.415	.239	.466	.312	.180
Margin(.05)	.236	.398	.235	.421	.249	.208
Margin(.1)	.083	.321	.153	.371	.183	.182

表 8 中国語-英語のセグメントレベルのピアソンの相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.515	.533	.350	.484	.494	.310
OneClass	.509	.528	.339	.474	.492	.287
Margin(.05)	.500	.523	.336	.461	.482	.274
Margin(.1)	.484	.503	.246	.429	.492	.293

表 6，7，8 に各言語対のセグメントレベルのピアソンの相関係数の結果を，表 9，10，11 に各言語対のセグメントレベルのスピアマンの順位相関係数の結果を示す．また，表 12 にシステムレベルのピアソンの相関係数を示す．システムレベルについては，ピアソンの相関係数の値のみを報告するが，これは WMT22 でも同様にシステムレベルではピアソン相関係数が報告されている．なお，システムレベルでは，WMT22 の全体の結果と TedTalks の結果のみを示す．

表 9 英語-ドイツ語のセグメントレベルのスピアマンの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.311	.448	.160	.381	.352	.297
OneClass	.308	.440	.177	.367	.333	.294
Margin(.05)	.307	.438	.174	.386	.298	.287
Margin(.1)	.194	.433	.132	.292	.256	.205

表 10 英語-ロシア語のセグメントレベルのスピアマンの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.262	.416	.208	.395	.328	.229
OneClass	.252	.406	.205	.396	.290	.224
Margin(.05)	.221	.391	.225	.361	.221	.207
Margin(.1)	.042	.350	.130	.300	.147	.128

表 11 中国語-英語のセグメントレベルのスピアマンの順位相関係数．

	wmt22					wmt21
モデル	mixed	news	conv.	ec.	social	tedtalks
MCD	.459	.494	.337	.462	.422	.302
OneClass	.455	.488	.321	.455	.422	.289
Margin(.05)	.443	.483	.299	.444	.405	.245
Margin(.1)	.424	.459	.230	.405	.407	.276

表 12 システムレベルのピアソンの相関係数．

	英語-ドイツ語		英語-ロシア語		中国語-英語	
モデル	mixed	tedtalks	mixed	tedtalks	mixed	tedtalks
MCD	.701	.662	.506	.823	.607	-.247
OneClass	.746	.634	.558	.857	.645	-.252
Margin(.05)	.751	.647	.539	.851	.631	-.295
Margin(.1)	.675	.531	.400	.790	.597	-.513