翻訳間で対応する多様な言語単位対の自動抽出：スパン予測に基づく手法の検討

本田友乃 [1]　藤田篤 [2]　宮田玲 [1]　影浦峡 [1]
[1] 東京大学大学院教育学研究科　[2] 情報通信研究機構
tomono20@g.ecc.u-tokyo.ac.jp

概要

産業翻訳や翻訳教育，翻訳評価における課題を背景として，同一の原文に対する異なる翻訳間の差異を記述するスキームが構築された [1, 2]．このスキームにより，人手による差異の分析は可能となったが，多様な翻訳対の分析には効率化が必要である．この課題に対し，先行研究 [3] では，スキームに基づく記述の工程の一部を対象に，翻訳間で対応する単位対の自動抽出としてタスクを定義し，実装したものの，精度が十分ではなかった．そこで，本研究では，スパン予測に基づく手法を検討した．

1　はじめに

産業翻訳や翻訳教育において，同一の原文に対する異なる翻訳を比較，分析することは重要である．翻訳教育では，自身の翻訳に対して，複数の翻訳候補の中から特定の言語表現を選択した理由を説明できることが重要であるとされている [4]．産業翻訳では，主にポストエディット (MT+PE) の形で機械翻訳や生成 AI の導入が進んでいるものの，人手翻訳 (HT) とこれらのシステムによる自動翻訳との間には決定的な差異があることが主張されている [5]．そのため，自動翻訳の出力の性質や，HT と比較した際の自動翻訳の限界点を記述的に解明することが重要である．また，翻訳評価では，複数の翻訳を比較する相対評価だけでなく，単一評価でも，参照訳を用いた評価 [6, 7, 8] をはじめ，翻訳の比較を介した評価を行うことがある．

以上を背景として，英日翻訳を対象に，同一の原文に対する異なる翻訳間の差異を記述するためのスキーム (差異分析スキーム) が構築された [1, 2]．差異分析スキームによる分析は，翻訳間で対応する単位対を同定する「分割」と，分割で得られた個々の単位対に対して差異カテゴリを付与する「分類」の

STの言語単位	TT_1の言語単位	TT_2の言語単位
You are required to pay the required fee.	必要な手数料を納付しなければならない。	所要の料金の支払いが求められます。
	⇩ 妥当な言語単位の対を階層的かつ網羅的に抽出	
the required fee	必要な手数料	所要の料金
required	必要な	所要の
fee	手数料	料金
You are required to pay	納付しなければならない	支払いが求められます
.	。	。

図 1　差異分析スキームに基づく単位対の抽出例

2 つの工程からなる．

差異分析スキームの構築により，翻訳間の差異の人手による分析は可能となった．一方で，作業負担を考慮すると，スキームを多数の翻訳対に適用することは困難である．この課題に対し，本田・藤田 [3] では，差異分析スキームにおける分割工程の自動化に向けて，翻訳間で対応する言語単位対の自動抽出としてタスクを定義し，実装したが，分析の効率化に十分な性能ではなかった．そこで，本研究では，スパン予測に基づくアラインメント手法を検討し，自動抽出タスクの精度向上に取り組んだ．

2　タスクの定義と先行研究

翻訳の原文 (Source Text) を ST，当該 ST に対する異なる翻訳テキスト (Target Text) を各々 TT_1，TT_2 と記す．差異分析スキーム [2] では，分割の工程向けに，分割・抽出の対象となる 9 種類の言語単位 (分析単位) と分割手順が定められている．本タスクでは，これらの手続きに従い，分割元の単位対を入れ子のない要素に分割する[1)]という制約のもと，所与の TT 対から分析単位対へと階層的に分割することで，図 1 のように，TT 対に含まれる単位対を網羅的に抽出する．なお，分析対象はあくまで TT 間の差異であ

1)　「入れ子のない要素」に分割する操作とは，分割元の分析単位の統語的な構造に従って，(A) 分割後の要素同士に重なりがないように分割する，(B) できるだけ大きな単位を抽出する，という 2 つの条件を満たす操作を表す．

るため，入出力における ST の単位は任意である．

先行研究 [3] では，語の対から大きい単位対をボトムアップに組み合わせるという手法を用いているが，分析の効率化には精度が十分ではなかった．

また，本タスクと類似する言語単位の対応づけに関する研究は，単語アラインメント [9, 10, 11, 12] やフレーズアラインメント [13] として行われてきた．しかし，これらの技術は，対象となる単位が単語やフレーズなどの特定の言語単位に固定されているという点で，本タスクと性質が異なる．

3 提案手法

本手法では，スパン予測に基づく単語アラインメントツールである，WSPAlign [10] の枠組みを適用する．WSPAlign [10] では，SQuAD [14] の実装が用いられており，特殊記号 ¶ で囲まれた原文中の単語をクエリとして設定すると，クエリに対応する翻訳文 (コンテクスト) 中の単語列がアンサーとして予測される．この形式を，原文と翻訳文を入れ替えた二方向に適用し，対称化することで，単語対が得られる．本タスクでは，抽出すべき単位対が多様であるため，スパン予測の手法が適すると判断した．また，訓練用データは，少量であれば著者自身で作成できるため，手法の適用が可能である．

単位対の自動抽出は，以下の手続きで行う．

1. TT 間の多様な単位対の取得
2. 単位対の追加
3. 後処理 (フィルタリング)

1 では，ファインチューニングされたモデルを用いて，TT 間の多様な単位対を取得する．テキストからすべての単語 n-gram ($1 \leq n \leq$ テキスト長) をクエリに用いる候補として抽出した上で，句構造解析器 [15] による統語構造を参照しつつ，文法的に妥当ではない単位を自動的に除外した．アラインメントの対称化は，Wu ら [10] と同様の方法で行ったが，対称化の閾値は，開発用データに対する F_1 値が最も高い値をモデルごとに採用した．1 で用いるモデルのファインチューニングに関しては，4 節で後述する．

2 と 3 の処理は，差異分析スキームの要件と開発用データによる出力を参照しながら，ヒューリスティックに定めた．2 では，1 の単位対に対して，(2-1) 単語アラインメント，(2-2) TT 間で表層上一致する単語対，(2-3) 1 の対称化で漏れた，一方が「対応なし」となる単位対 (null alignment) を追加する[2)]．

3 では，2 までで得られた単位対から，(3-1) トークン数の比が 4 を超える単位対，(3-2) 分割元が同じ単位対，(3-3) 文法的に妥当ではない単位対を除外する[3)4)]．

なお，入力となる TT 対に対しては，先行研究 [3] と同様に，以下の前処理を行った．

- テキスト間の文単位での対応づけ
- メールアドレスと URL の抽象化
- 分かち書き[5)]

4 多様な単位対の予測に関する実験

提案手法の手続き 1 で用いるモデルについて，複数の事前学習済みモデルを対象に，人手で作成した訓練用データ (ゴールドデータ) と自動的に作成した訓練用データ (シルバーデータ) によるファインチューニングを行い，提案手法を評価した．

4.1 実験に用いたデータ

実験には，MultiEnJa[6)] の HT と MT+PE の対のうち，1 種類を開発用データ (文書 284)，それ以外の 2 種類 (文書 373，088) を評価用データとして用いた．開発用データと文書 373 は先行研究 [3] と同じデータ，文書 088 は本研究で新たに作成したデータである．

ゴールドデータは，MultiEnJa から，開発，評価用データを除いて無作為に選択した 6 種類の HT と MT+PE の対から作成した．シルバーデータには，未使用の MultiEnJa の HT と MT+PE の対に加えて，ParaNatCom [20] の HT の対，Asian Language Treebank (ALT) の対訳データ [21] に含まれる HT と ALT に対応する MT+PE である staged-PE [22] の対を用いた．データ量は付録 A を参照されたい．

2) (2-1) では，英日の単語アラインメント用にファインチューニングされた WSPAlign-ft-kftt [10] のモデル，(2-2) では Meteor [16] を利用した．ただし，(2-2) で，ひらがな一文字同士の単語対は，不適切な助詞の対が多いため，追加の対象から除いた．また，(2-3) で，一方の単位がひらがな一文字の場合または文法的に妥当ではない単位の場合は，追加の対象から除いた．

3) (3-2) は，TT 間で表層上一致する単位対は分割を停止するという差異分析スキームの要件を実装した処理である．なお，(3-3) には句構造解析器 [15] を用いた．

4) 提案手法の手続き 1 で得られる文字列スパンと手続き 2, 3 の処理単位である形態素の対応付けに Jigg [17] を用いた．

5) ST に対しては NLTK [18] を，TT に対しては MeCab 及び IPA 辞書 [19] を使用して分かち書きを行った．ただし，分かち書きの後，正解データと照らし合わせて正解データと単位が異なる箇所は，人手による修正を施した．

6) https://github.com/tntc-project/MultiEnJa

表 1 実験結果 (%): gold，silver はそれぞれゴールドデータとシルバーデータを表す．

用途	文書	事前学習済みモデル	訓練の形式 (手法)	閾値	TP	TP+FN	TP+FP	再現率	適合率	F_1 値
開発	284	WSPAlign-mbert-base	-	0.5	191	328	275	58.2	69.5	63.4
開発	284	WSPAlign-mbert-base	gold (en・ja)	0.5	210	328	314	64.0	66.9	65.4
開発	284	bert-base-multilingual-cased	gold (en・ja)	0.5	214	328	327	65.2	65.4	65.3
開発	284	bert-base-japanese-v3	gold (en・ja)	0.7	239	328	344	**72.9**	69.5	**71.2**
開発	284	bert-large-japanese-v2	gold (en・ja)	0.8	237	328	356	72.3	66.6	69.3
開発	284	roberta-large-japanese-char-wwm	gold (en・ja)	0.9	232	328	335	70.7	69.3	70.0
開発	284	WSPAlign-mbert-base	gold (ja)	0.5	205	328	300	62.5	68.3	65.3
開発	284	bert-base-multilingual-cased	gold (ja)	0.5	207	328	300	63.1	69.0	65.9
開発	284	bert-base-japanese-v3	gold (ja)	0.5	232	328	343	70.7	67.6	69.1
開発	284	bert-large-japanese-v2	gold (ja)	0.5	222	328	324	67.7	68.5	68.1
開発	284	roberta-large-japanese-char-wwm	gold (ja)	0.8	228	328	329	69.5	69.3	69.4
開発	284	WSPAlign-mbert-base	gold (ja + ST)	0.5	208	328	305	63.4	68.2	65.7
開発	284	bert-base-multilingual-cased	gold (ja + ST)	0.5	203	328	299	61.9	67.9	64.8
開発	284	bert-base-japanese-v3	gold (ja + ST)	0.5	222	328	325	67.7	68.3	68.0
開発	284	bert-large-japanese-v2	gold (ja + ST)	0.9	236	328	339	72.0	**69.6**	70.8
開発	284	roberta-large-japanese-char-wwm	gold (ja + ST)	0.9	230	328	346	70.1	66.5	68.3
開発	284	bert-base-japanese-v3	silver	0.5	228	328	349	69.5	65.3	67.3
開発	284	bert-base-japanese-v3	silver → gold	0.7	231	328	351	70.4	65.8	68.0
開発	284	bert-base-japanese-v3	silver → silver + gold	0.7	232	328	354	70.7	65.5	68.0
開発	284	bert-base-japanese-v3	silver → silver + gold → gold	0.7	232	328	360	70.7	64.4	67.4
開発	284	bert-base-japanese-v3	silver + gold	0.7	238	328	356	72.6	66.9	69.6
開発	284	bert-base-japanese-v3	silver + gold → gold	0.7	239	328	358	**72.9**	66.8	69.7
開発	284	-	ボトムアップ手法 [3]	-	214	328	310	65.2	69.0	67.1
評価	373	bert-base-japanese-v3	gold (en・ja)	0.7	296	434	459	68.2	64.5	66.3
評価	373	-	ボトムアップ手法 [3]	-	245	434	412	56.5	59.5	57.9
評価	088	bert-base-japanese-v3	gold (en・ja)	0.7	277	470	510	58.9	54.3	56.5

ゴールドデータと開発用データは，差異分析スキームを用いた分割経験のある著者 1 名が作成し，別の著者によるチェックを行った．評価用データについては，複数名による差異分析スキームの評価実験 [2] の結果を参照して正解を定めた．

4.2 ファインチューニング

ゴールドデータによるファインチューニングでは，データの形式として，(1-a) ST-TT_1，ST-TT_2，TT_1-TT_2 の 3 種類の対 (6 方向) を含むデータ，(1-b) TT_1-TT_2 の対 (2 方向) のデータ，(1-c) TT_1-TT_2 の対 (2 方向) の各方向に対し，クエリに対応する ST の文を [SEP] トークンで連結させたデータ，の 3 種類を試した．(1-a) は英日の対も用いてデータ量を増やす試み，(1-b) は自動抽出の対象言語である日本語の対のみを利用する試み，(1-c) は TT に共通する ST の情報を同時に参照する試みである．

また，ファインチューニングに用いる事前学習済みモデルとして，単語アラインメントのデータで学習済みの WSPAlign-mbert-base[7] [10] とそのベースモデルである mBERT[8] [23]，及び，日本語データで訓練された 2 種類の BERT [23] モデル[9] を選択した．さらに，スパン予測との相性を確認するために，文字単位でのトークナイズを行う RoBERTa [24] の日本語モデル[10] も評価した[11]．

ゴールドデータのみを用いた評価実験を実施後，訓練用データの分量を増やすため，ゴールドデータによるファインチューニングにおいて再現率が最高のモデルを用いてシルバーデータを自動的に作成した[12]．シルバーデータによるファインチューニング

7) https://huggingface.co/qiyuw/WSPAlign-mbert-base
8) https://huggingface.co/google-bert/bert-base-multilingual-cased
9) https://huggingface.co/tohoku-nlp/bert-base-japanese-v3, https://huggingface.co/tohoku-nlp/bert-large-japanese-v2
10) https://huggingface.co/ku-nlp/roberta-large-japanese-char-wwm
11) WSPAlign-ft-kftt [10]，xlm-roberta-large [25]，infoxlm-large [26] は，他のモデルを一定程度下回る性能であった．
12) データの品質を保つため，閾値として，適合率が最も高い値を用いた．選択された閾値は 0.7 であり，結果として，

では，ゴールドデータとの組み合わせや訓練の順序が異なる 6 種類のカリキュラムを試した．

4.3 実験結果

自動抽出の結果は，先行研究 [3] と同様に，正解データとの再現率，適合率，F_1 値で評価した．実験結果を，先行研究 [3] で開発用データにおける F_1 値が最も高かった結果と併せて，表 1 に示す．また，訓練による効果の多寡を示すため，参考として，WSPAlign-mbert-base [10] をファインチューニングなしで適用した結果も記載した．評価用データに関しては，開発用データで F_1 値が最も高いモデルを適用した結果のみを記載した．

開発用データに対しては，英日の 3 種類の対を含むゴールドデータで訓練した bert-base-japanese-v3 のモデルが，F_1 値で 71.2%と最も高い性能となった．また，全体的な傾向として，品質の劣るシルバーデータを大量に用いるよりも，少量の高品質なゴールドデータを用いた方が性能が高いこと，日本語モデルで比較的高い性能が得られることがわかった．

評価用データに対しては，先行研究 [3] と比較可能な文書 373 において，先行研究から F_1 値で 8 ポイント以上高い性能が得られた．一方で，開発用データと比較すると，文書 373，088 で各々 4.9，14.7 ポイント F_1 値が低下した．

4.4 エラー分析

自動抽出で得られた単位対の特徴を分析するために，先行研究 [3] と同様に，評価用データに対するエラー分析を行った．より詳しい分析は，付録 C を参照されたい．

4.4.1 抽出もれ

評価用データにおける分析単位対の種類別の再現率を表 2 に示す．表 2 より，語と語，記号と記号などの小さい単位対は比較的多く抽出できているのに対し，語句相当と語句相当[13)]，節と節，句と句などの大きい単位対はあまり抽出できていないことがわかる．これらの傾向は先行研究 [3] と同じである．一方で，先行研究 [3] の手法では，null alignment が原理的に抽出できないという課題があったのに対し，本手法では一定程度抽出可能となった．

F_1 値を基準とした閾値と一致した．

13） 語句相当の単位には，括弧類を含む全体の要素，専門用語，固有表現が含まれる．

表 2 評価用データにおける分析単位対の種類別の再現率

HT \ MT+PE	節	語句相当	句	複合表現	語	記号	なし
節	21/42	-	0/3	-	-	-	-
語句相当	-	24/65	0/1	-	-	-	-
句	1/4	-	54/114	4/16	1/3	-	-
複合表現	-	0/1	9/15	82/131	12/24	-	3/6
語	-	1/1	0/3	8/17	238/294	3/3	3/15
記号	-	-	-	-	-	88/95	4/16
なし	-	-	-	3/8	5/14	9/13	-

4.4.2 抽出誤り

誤って抽出された対を，(i) 単位・対ともに適切だが不要な対，(ii) 単位は適切だが対として適切でない対，(iii) 一方または双方の単位が適切でない対，の 3 種類の誤りに分類した．表 3 より，いずれの文書でも，(i) が 7 割程度，(ii) が 3 割弱，その他が (iii) という結果になった．先行研究 [3] では，文書 373 に関して，(ii) が 134 件で誤り全体の過半数を占めており，(iii) が 24 件含まれていたことをふまえると，本手法によって，意味的に対応する妥当な単位対を抽出する性能が向上したとわかる．

表 3 抽出誤りの内訳

誤りの種類	373	088
(i) 単位・対ともに適切だが不要な対	114	162
(ii) 単位は適切だが対として適切でない対	46	56
(iii) 一方または双方の単位が適切でない対	3	15

5 おわりに

本稿では，翻訳間の差異の記述の効率化を主な目的として，翻訳間で対応する多様な言語単位対の自動抽出に取り組み，スパン予測による単語アラインメント [10] を応用する手法について検討した．実験の結果，先行研究 [3] と比較可能な評価用データに関しては，F_1 値を 8 ポイント以上改善した．また，エラー分析の結果は，意味的に対応する単位対の抽出性能が先行研究 [3] から向上したことを示している．

今後は，差異分析スキームを構築した背景と目的に立ち戻り，人間による分割結果との比較も含めて自動抽出の出力を分析し，自動抽出の限界点についても検討したい．特に，翻訳評価尺度が近年の自動翻訳の高い性能に必ずしも対応しきれていないという課題 [27] を考慮すると，本研究のような精密な言語操作を要するタスクを通して，翻訳における言語単位の扱いを診断することは，間接的ではありながらも重要な意義を有すると考えている．

謝辞

本研究の成果の一部は，UTokyo Azure [28] を利用して得られた．また，本研究は JSPS 科研費 JP24KJ0850，JP23K28378 の助成を受けた．

参考文献

[1] 本田友乃, 山本真佑花, 影浦峡. 異なる翻訳間の差異を記述するためのスキームの構築. 通訳翻訳研究への招待, No. 24, pp. 1–21, 2022.

[2] 本田友乃, 藤田篤, 山本真佑花, 影浦峡. 異なる翻訳間の差異を記述するためのスキーム—体系の洗練とメタ言語としての評価—. 通訳翻訳研究, Vol. 23, pp. 83–103, 2023.

[3] 本田友乃, 藤田篤. 同一の原文書に対する複数の翻訳文書間で対応する言語単位対の自動抽出. 言語処理学会第 30 回年次大会 発表論文集, pp. 1797–1802, 2024.

[4] European Master's in Translation. European Master's in Translation competence framework 2022, 2022.

[5] International Federation of Translators. FIT position paper on machine translation in the age of AI, 2025.

[6] 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.

[7] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In **Proc. of AMTA (Technical Papers)**, pp. 223–231, 2006.

[8] Maja Popović. chrF: Character n-gram F-score for automatic MT evaluation. In **Proc. of WMT**, pp. 392–395, 2015.

[9] Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced optimal transport for unbalanced word alignment. In **Proc. of ACL (Volume 1: Long Papers)**, pp. 3966–3986, 2023.

[10] Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In **Proc. of ACL (Volume 1: Long Papers)**, pp. 11084–11099, 2023.

[11] Gaetan Latouche, Marc-André Carbonneau, and Benjamin Swanson. BinaryAlign: Word alignment as binary sequence labeling. In **Proc. of ACL (Volume 1: Long Papers)**, pp. 10277–10288, 2024.

[12] Zhongtao Miao, Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. Improving word alignment using semi-supervised learning. In **Findings of ACL**, pp. 19871–19888. Association for Computational Linguistics, 2025.

[13] Yuki Arase and Jun'ichi Tsujii. Compositional phrase alignment and beyond. In **Proc. of EMNLP**, pp. 1611–1623, 2020.

[14] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proc. of EMNLP**, pp. 2383–2392, 2016.

[15] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In **Proc. of ACL (Volume 1: Long Papers)**, pp. 277–287, 2017.

[16] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In **Proc. of WMT**, pp. 376–380, 2014.

[17] Hiroshi Noji and Yusuke Miyao. Jigg: A framework for an easy natural language processing pipeline. In **Proc. of ACL (System Demonstrations)**, pp. 103–108, 2016.

[18] Steven Bird, Ewan Klein, and Edward Loper. **Natural language processing with Python**. O'Reilly Media, 2009.

[19] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proc. of EMNLP**, pp. 230–237, 2004.

[20] Masao Utiyama. ParaNatCom—Parallel English-Japanese abstract corpus made from Nature Communications articles, 2020.

[21] Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. Introduction of the Asian Language Treebank. In **Proc. of O-COCOSDA**, pp. 1–6, 2016.

[22] Atsushi Fujita. Attainable text-to-text machine translation vs. translation: Issues beyond linguistic processing. In **Proc. of MT Summit (Research Track)**, pp. 215–230, 2021.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. of NAACL-HLT (Volume 1: Long and Short Papers)**, pp. 4171–4186, 2019.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. **arXiv:1907.11692**, 2019.

[25] 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 **Proc. of ACL**, pp. 8440–8451, 2020.

[26] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In **Proc. of NAACL-HLT**, pp. 3576–3588, 2021.

[27] Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In **Proc. of WMT**, pp. 436–483, 2025.

[28] 中村誠, 中村遼, 有馬和美, 佐々木馨, 山肩洋子, 伊藤研一郎, 西島学, 宮嵜洋, 石崎勉, 玉造潤史, 雨宮智浩, 千葉滋, 田浦健次朗. UTokyo Azure: クラウドコンピューティングサービスの学内展開. 学術情報処理研究, Vol. 29, No. 1, pp. 180–186, 2025.

A　実験に用いたデータ

4.1 節で述べた実験データにおけるクエリの数を表 4 に示す．ゴールドデータには，文書 002，032，117，198，407，497 の 6 種類が含まれる．

表 4　実験データにおけるクエリの数

用途	データ	クエリの数 (en・ja)	クエリの数 (ja)
訓練	ゴールド	14,865	5,192
	シルバー	-	145,035
開発	文書 284	-	1,112
評価	文書 373	-	1,394
	文書 088	-	1,720

B　実験設定

4 節のファインチューニングでは，開発用データを用いて，表 5 に示すハイパーパラメーターを設定し，損失関数によるモデル選択を行った．なお，ゴールドデータの学習率に関しては，WSPAlign-mbert-base [10] でのみ 3e-6 を用いた．また，GPU には，Azure NC A100 v4 仮想マシン上の NVIDIA A100 PCIe GPU (80 GB) を用いた．

表 5　ハイパーパラメーターの設定

パラメーター	ゴールドデータ	シルバーデータ
学習率	1e-5	3e-6
バッチサイズ	16	16
累積勾配のステップ	4	16
エポック数	30	3

C　エラー分析

C.1　抽出もれ

分析単位対の種類別の再現率を，文書ごとに表 6，7 に示す．

4.4.1 節で，語句相当と語句相当，節と節，句と句などの大きい単位対の抽出もれが比較的多いことを述べた．語句相当と語句相当の単位対には，専門用語や固有表現の対が多く，「『リスク評価リスク軽減戦略追加 (REMS Supplement)』」と「『REMS Supplement』」のように，単位間で構造や表記が大きく異なる対の抽出もれが目立った．

節や句の抽出もれに関して，例えば，以下の句の対では，四角括弧のような分割を行い，節と句の対，複合表現 (複合名詞) と複合表現の対の 2 種類を抽出すべきである．

- 【HT】[eCTD フォーマットの DMF Type lll を提出する] [遵守日]
- 【MTPE】[DMF Type lll の eCTD 形式による提出] の [遵守期限日]

しかし，自動抽出では，上記の節と句の対は抽出せず，以下のような小さい単位対を一気に抽出していた．

- 【HT】[eCTD フォーマット] [の] [DMF Type lll] を [提出する] [遵守日]
- 【MTPE】 [DMF Type lll] [の] [eCTD 形式] による [提出] の [遵守期限日]

表 6　分析単位対の種類別の再現率 (文書 373)

HT \ MT+PE	節	語句相当	句	複合表現	語	記号	なし
節	7/17	-	0/1	-	-	-	-
語句相当	-	5/12	-	-	-	-	-
句	0/2	-	20/40	2/9	0/1	-	-
複合表現	-	-	6/10	49/78	8/17	-	1/3
語	-	-	0/3	3/6	136/161	1/1	1/5
記号	-	-	-	-	-	43/47	2/5
なし	-	-	-	2/3	3/5	7/8	-

表 7　分析単位対の種類別の再現率 (文書 088)

HT \ MT+PE	節	語句相当	句	複合表現	語	記号	なし
節	14/25	-	0/2	-	-	-	-
語句相当	-	19/53	0/1	-	-	-	-
句	1/2	-	34/74	2/7	1/2	-	-
複合表現	-	0/1	3/5	33/53	4/7	-	2/3
語	-	1/1	-	5/11	102/133	2/2	2/10
記号	-	-	-	-	-	45/48	2/11
なし	-	-	-	1/5	2/9	2/5	-

C.2　抽出誤り

3 種類の抽出誤りのうち，(iii) 一方または双方の単位が適切でない対には，「eCTD 提出基準 (PDF - 91KB)」の赤字部分のように，括弧類の一部が含まれる単位など，1 つの独立した言語単位として認められない事例が見られた．

(ii) 単位は適切だが対として適切でない対の誤りには，小さい単位対が多く，表層上一致するが文における役割が異なる助詞など，機能語に関する誤りが目立った．

(i) 単位・対ともに適切だが不要な対には，大きい単位対と小さい単位対の双方が含まれていた．小さい単位対の抽出誤りには，差異分析スキームにおける，述語表現 (分析単位では複合表現) を抽出した段階で分割を止めるという分割停止条件に違反して，述語表現の対からさらに細かい語や形態素の単位対を抽出する事例などが見られた．

大きい単位対の抽出誤りに関して，例えば，以下の文対を差異分析スキームの分割手続きに従って一段階分割した結果は，四角括弧のようになる．

- 【HT】 [2011 年建築工事（登録）法] では [、] [登録区分にある建築工事請負業者が規定の財務要件を満たしていることを Board が確認すること] が [求められています] [。]
- 【MTPE】 [2011 年建築サービス (登録) 法] は [、] [登録クラスの建築サービス請負業者が所定の財務要件を満たしていることを委員会が確認すること] を [求めています] [。]

しかし，自動抽出では，以下の赤字部分のように，文全体における機能語の役割を考慮せずに，複数の単位をまとめて抽出する事例が目立った．

- 【HT】[2011 年建築工事（登録）法] では [、] [登録区分にある建築工事請負業者が規定の財務要件を満たしていることを Board が確認することが求められています] [。]
- 【MTPE】 [2011 年建築サービス (登録) 法] は [、] [登録クラスの建築サービス請負業者が所定の財務要件を満たしていることを委員会が確認することを求めています] [。]