特許請求項翻訳における
単語対応に基づく節分割モデルの有効性

西村柾人[1] 宇津呂武仁[1] 永田昌明[2]
[1]筑波大学大学院 システム情報工学研究群 [2]NTT コミュニケーション科学基礎研究所
s2320779_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

概要

特許請求項は特許範囲を規定する重要部分だが，その長さや独特の書式が原因で NMT モデルでは訳抜けや繰り返しといった誤訳を引き起こしやすい．本論文では，この課題を解決するために，節分割モデルを用いた分割統治翻訳手法を提案する．翻訳元文を節分割モデルを用いて節に分割し，それぞれの節を節翻訳モデルで翻訳，その後並び替え・編集モデルで最終的な翻訳文を生成する．さらに，節分割モデルと節翻訳モデルを訓練するための節単位の対訳コーパスを，単語対応情報をもとに文単位の対訳コーパスから作成する手法を提案する．実験では，提案手法が通常のモデルを BLEU で上回り，訳抜けや繰り返しの改善を確認した．

1 はじめに

特許文書における特許請求項は，特許の権利範囲を定義する上で極めて重要な部分である．しかし，その文の長さや独特の記述形式のため，ニューラル機械翻訳 (NMT) モデルを用いた翻訳では，訳抜けや繰り返しといった問題が発生しやすい．長文翻訳におけるこれらの問題に対して加納ら [3] は，英日翻訳において，入力文を構文解析に基づいた節に分割して翻訳後に再構成する「分割統治ニューラル機械翻訳」手法を提案した．ここで，この手法は翻訳精度向上の可能性を示したものの，その一方で，適切な節の単位の選定や，節翻訳後の再構成精度に課題が残っていた．

これに対し石川ら [1] は，文献 [3] で課題とされていた，節分割単位の課題と節分割後の節の翻訳精度の二つの課題に対して，接続詞による節分割の採用と，節翻訳モデルと並び替え・編集モデルの 2 つに事前学習済みモデルの mBART [5] を用いることにより翻訳精度の向上を試みた．それに加えて，節翻訳モデルを節単位の疑似対訳データでファインチューニングすることによって，節の翻訳精度改善を試みた．実験では，過剰な長さの訳出が大きく減少し，ハルシネーションや繰り返しの抑制を確認している．このように，分割統治翻訳手法を用いることで，長文翻訳において従来の NMT モデルが引き起こしやすい翻訳ミスを抑制できる可能性が示唆されている．しかし，文献 [1] における課題として，節の分割単位が接続詞のため適切に短い節に分割できない長文が存在する点と，節翻訳モデル訓練時に擬似対訳データを用いており，実際に収集された実対訳データを用いていない点が挙げられる．

以上をふまえて，本論文では，特に特許請求項日英翻訳を対象とする分割統治翻訳手法において，先行研究 [3, 1] とは異なるアプローチを提案する．本論文の手法においては，まず，翻訳元の特許請求項文を翻訳モデルが翻訳しやすい単位の節に分割する節分割モデルを提案する．特に，本論文では，対訳文中の単語対応に基づいて設定された節単位が二言語間で整合することを条件とすることにより，最終的な翻訳文で訳抜けや繰り返しといった誤訳の発生を抑え，より正確な翻訳文を生成することを実現した．具体的な手法としては，単語アライメントツールを用いて，元の対訳コーパスから節単位の対訳データを高品質に生成する手法，および，作成した節単位の対訳コーパスを用いて以下の 3 つのモデルを構築する手法を提案する：

1. 翻訳元の日本語文を節単位に分割する節分割モデル．
2. 節翻訳に特化した節翻訳モデル．
3. 翻訳された節を並び替え・編集し，最終的な翻訳文を生成する並び替え・編集モデル．

実験として，これらのモデルを統合した翻訳手

法を，日英特許対訳コーパス JaParaPat [6] を用いて評価した．その結果，特許請求項の日英翻訳において，提案手法は通常の NMT モデルと比較して，BLEU において統計的に有意な改善を示した．また，通常モデルによる翻訳結果と比較して，提案手法において訳抜けの発生を抑制できていることを確認し，より正確な翻訳結果を得られることが示された．

2 関連研究

長文翻訳の課題に対しては，従来からさまざまなアプローチが検討されてきた．Sudoh ら [11] は，長い文を翻訳するため，統計的機械翻訳に分割統治的翻訳手法を用いた．入力文を構文解析に基づいた節の単位に区切って翻訳し，その結果を節の階層構造に基づいて並べ替えることで，精度向上を実現させた．

NMT において，長文を節に分割し，分割された節ごとに翻訳した後に，前から順に結合し直す手法として Pouget-Abadie ら [10] の自動分割の手法がある．この手法は RNN を用いて，長文を分割して翻訳する際の最適な位置を予測し，モデルが翻訳し易いように入力文を分割する手法である．しかし，この手法は英仏翻訳を対象としており，本論文で行う日英翻訳のような語順が大きく異なる言語対においては，翻訳後の再構成において，不自然な語順が生じやすいという課題が残されていた，この課題に対し，加納ら [3] は，英日翻訳において，長文を分割して翻訳した後にそれぞれの節を適切な順序に並び替えるニューラルネットワークモデルを作成した．さらに石川ら [2] は，等位接続詞を基準とした新しい英文の節の分割単位及び, 分割後の節翻訳において文内コンテキストを参照して翻訳を行うモデルの学習手法を提案し，長文の英日翻訳において翻訳精度向上を達成した．

3 提案手法

図 1 に，提案手法全体の構造を示す．翻訳元の日本語特許請求項文は，まず節分割モデルによって複数の節に分割される．その後，各節が節翻訳モデルによって翻訳され，最後に並び替え・編集モデルによって統合され，翻訳後の英語特許請求項が生成される．本手法により，従来の NMT モデルで発生しやすい訳抜け・繰り返しを抑制することを目指す．

3.1 節単位の対訳コーパス

本論文では，Zhang ら [15] の長文の対訳データから対訳部分文を生成する手法を参考に，節単位の対訳コーパスを自動的に生成する手法を提案する．この手法では，文単位の対訳コーパスに対して，単語アライメントツールの WSPAlign [14] を用いて単語対応情報を取得する．単語対応情報を基に，文中の対応する節を抽出し，節単位の対訳データを生成する．節単位の対訳コーパスは以下のような手順で作成する．本論文では文献 [15] の報告を基に，単語含有割合の閾値を 0.5 と設定した．

1. 特許対訳データの対訳文に対して WSPAlign を使用することで単語対応を取得する．
2. 日英の対訳文それぞれに対して，「、」「,」「。」「.」「;」「:」などの区切り記号の位置で複数の節に分割する．
3. 単語対応情報から，対訳節ごとの単語含有割合を計算し，割合が 0.5 を超えた場合，対象の節は対応関係があると判定する．
4. 節対応が 1 対多, 多対多となる場合は，1 対 1 になるように複数節側を 1 つにまとめる．

上記の手順を特許対訳コーパスから抽出した対訳文に適用することで節対訳コーパスを作成する

3.2 節分割モデル

本論文では，Wicks と Post [13] が提案した文分割モデル ERSATZ を節分割モデルの基盤とし，このモデルを節単位の分割に適用するように訓練を行うことによって節分割モデルを作成した．ERSATZ は，文分割をバイナリ分類タスクとして定式化しており，文分割候補となる句点 (「。」や「.」など) に対して「文中」か「文末」かを予測する．本論文では，この文分割を節分割に拡張するために，節分割候補となる読点 (「、」や「,」など) を用いて節分割モデルを作成する[1)]．モデルの訓練には 3.1 節で提案した節単位の対訳コーパスを用いた．

訓練に使用するデータは，節対訳コーパス内の日本語節を文単位で抽出し，節の分割位置の句読点に文末ラベルを付与することで正解データを作成した．これにより，日本語特許請求項文に対して単語対応情報をもとにした節分割が行えるモデルを作成

1) 実際には，補足説明を表す () 中の文分割位置において節分割を行うために，読点 (「、」や「,」など) だけでなく句点 (「。」や「.」など) もあわせて用いる．

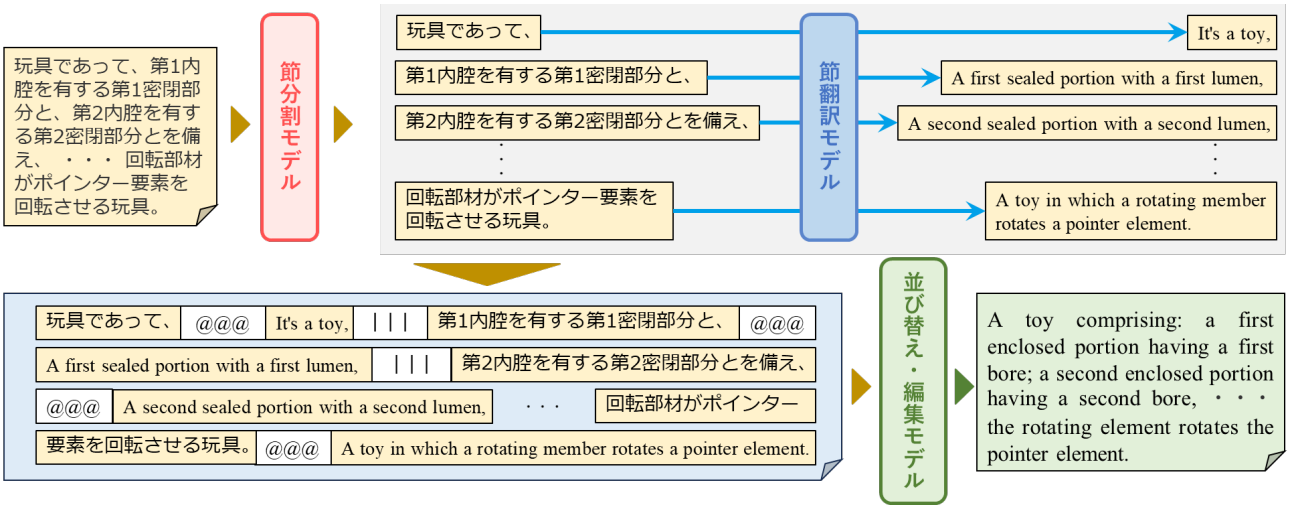

図 1　提案手法の流れ

した．

3.3　節翻訳モデル

本論文では，節翻訳に特化した節翻訳モデルを作成するために，事前に特許対訳コーパスから作成した日英翻訳モデルを，3.1 節の手法により作成された節単位の対訳コーパスでファインチューニングした．この節翻訳モデルは，分割された節を翻訳する際に分割によって失われた文脈を勝手に補おうとする事象を抑制し，より正確な翻訳を実現することを目的としている．

3.4　並び替え・編集モデル

並び替え・編集モデルを用いる理由としては，節翻訳モデルによって翻訳した複数の節を目的言語文である英語の一つの文に戻すためである．日本語と英語の語順は大きく異なることから，日本語文を節ごとに分割して接続するのみでは，日本語と英語間での語順の変化に対応することができない．そのため，節翻訳モデルによって得られた，節に分割された翻訳節を接続する際に並び替え・編集モデルを用いることで，適切な語順に並び替えることを期待している．

並び替え・編集モデルの訓練データは，コーパスの日本語文を節分割モデルで分割し，日本語の節とその翻訳結果の英語節を繋げたものを入力データ，コーパスの英語文を教師データとして利用する．モデルの語彙の中に特殊トークンとして “@@@” と “|||” を登録しておく．“@@@” は日本語の節とその翻訳結果の英語節を繋げるために用い，“|||” はそれらの節のペアを接続するために用いる．入力データをこのような形にしている理由としては，翻訳結果の英語節だけを入力データとして用いると，翻訳された各節どうしがどのような関係性をもつのかという情報が欠落してしまうため，日本語文の情報を付与するためである．

並べ替え・編集モデルの入力は日英両方の単語が含まれた文となるため，日英両方の理解力が必要となる．そのため並び替え・編集モデルは日英双方向の翻訳モデルを上記の手順で作成した訓練データでファインチューニングすることによって作成する．この事前訓練がモデルの性能に与える影響については，付録 B に示す．

4　実験

4.1　実験設定

本論文では，日英特許対訳コーパスの JaParaPat [6] を基に，日英翻訳の実験を行った．使用したデータは，2016 年から 2020 年までの特許全文の対訳データを訓練データ，2021 の前半までの特許請求項対訳データをテストデータとした．

機械翻訳ソフトウェアは fairseq [7] を使用し，ベースラインモデル, 節翻訳モデル, 並び替え・編集モデルのモデルアーキテクチャに Transformer big [12] を使用した．トークナイズには sentencepiece [4] を使用した．特許対訳データから 10M 文対をランダムにサンプリングしてモデルの訓練を行った．語彙数は日英ともに 32K とした．また，節分割モデルの訓練には ERSATZ[2] を利用した．

ベースラインモデルと提案手法の 3 種類のモデ

2)　https://github.com/rewicks/ersatz

表 1 提案手法およびベースラインモデルに使用したデータの概要

モデル	使用データ	データ数
ベースラインモデル	JaParaPat2016〜2020	61,364,685 文対
節分割モデル	節対訳コーパス (claims)	200,462 文
節翻訳モデル	JaParaPat2016〜2019 節対訳コーパス	49,474,547 文対 5,480,682 節対
並び替え・編集モデル	JaParaPat2016〜2020(双方向) JaParaPat2016〜2020(claims)	109,028,682 文対 2,613,107 文対

表 2 各評価対象の文数および BLEU

評価対象	文数	ベースライン	提案手法
全体	238,902	55.5	**56.6****
長文	33,959	50.1	**51.6****

ルの訓練に使用したデータの概要を表 1 に示す．節単位の対訳コーパスは，2020 年のデータの半数 (5,976,295 文対) に対して WSPAlign[3] を使用して単語対応情報を取得し，3.1 節の手法に沿って作成した．これによって 5,480,682 対の節対訳データを持つ節単位の対訳コーパスが作成された．節分割モデルの訓練には節単位の対訳コーパスから，特許請求項文を分割して作成した日本語節データを使用した．節翻訳モデルは，2016 年から 2019 年の特許全文の対訳データで事前訓練し，節単位の対訳コーパスの全データを使用してファインチューニングすることで作成した．並び替え・編集モデルは，2016 年から 2020 年の特許全文の双方向の対訳データで事前訓練し，2016 年から 2020 年の特許請求項の日本語文を，節分割モデルと節翻訳モデルを用いて 3.4 節の手法で作成した訓練データでファインチューニングして作成した．

評価指標として，BLEU [8] を使用した．BLEU は sacreBLEU[4] [9] を用いて計測した．特許翻訳では専門用語を正しく訳出できることが重視されるため，本論文では BLUE を主たる評価尺度とした．

4.2 結果

本論文では，2021 年の特許データから特許請求項のみを抽出したテストデータ（238,902 文）を用いて提案手法の性能を評価した．その結果，提案手法は BLEU において 56.6 を記録し，ベースラインモデルの 55.5 を統計的に有意に上回った（$p < 0.01$）．これにより，提案手法が全体的な翻訳精度を向上させることが確認された．

3） https://github.com/qiyuw/WSPAlign
4） https://github.com/mjpost/sacrebleu

また，テストセットの中で翻訳元日本語文のサブワードトークンが 100 トークン以上の長文のみを抽出したサブセットに対しても性能を評価した．結果としては，提案手法は 51.6 となり，ベースラインモデルの 50.1 を統計的に有意に上回った．また，テストセット全体では BLEU の上昇幅が 1.1 ポイントであったが，長文においては 1.5 ポイントの上昇が見られ，提案手法は特に長文において顕著な改善を示した．

さらに，提案手法が訳抜けや繰り返しといった翻訳ミスを抑制できるかを分析するため，ベースラインモデルおよび提案モデルの訳出文と参照訳文の文長比に着目し，その傾向を観察した．テストセット全体の分析結果では，訳抜けが発生する可能性が高い文の数は，提案手法で 515 文であるのに対し，ベースラインモデルでは 900 文であった．この結果から，提案手法はベースラインモデルに比べて訳抜けを効果的に抑制できていると考えられる．一方で，繰り返しに関しては，提案手法で発生した文が 376 文であるのに対し，ベースラインモデルでは 273 文であった。このことは，ベースラインモデルによる翻訳の方が繰り返しの発生が少ない可能性を示唆している．繰り返しの多くは，並べ替え・編集モデルで発生している可能性があり，並べ替え・編集モデルにおける繰り返し抑制を行う必要がある．これらの分類結果および詳細な分析については，付録 C に記載する．

また，実際の提案手法による訳抜け・繰り返しの翻訳改善例を表 3 に示す．図中の例のようにベースラインモデルで訳抜け・繰り返しが起きている文に対して，提案手法ではそれらの翻訳ミスが改善され，より正確な翻訳が行えることが確認できた．

5 おわりに

本論文では，特許請求項の日英翻訳における課題である「訳抜け」や「繰り返し」といった翻訳ミスを解決するため，特許請求項が長文や独特な構造を持つことに着目し，特許請求項をより翻訳しやすい単位に分割する節分割モデルを用いた翻訳手法の提案を行った．実験の結果，提案手法は BLEU においてベースラインモデルを統計的に有意に上回る性能を示し，訳抜けの発生を抑制できていることを確認した．これにより，提案手法が特許請求項の翻訳においてより正確な翻訳が可能であることが示された．

参考文献

[1] 石川隆太, 加納保昌, 須藤克仁, 中村哲. 事前学習モデルによる分割統治ニューラル機械翻訳. 言語処理学会第 29 回年次大会論文集, pp. 1451–1456, 2023.

[2] 石川隆太, 加納保昌, 須藤克仁, 中村哲. 文内コンテキストを利用した分割統治ニューラル機械翻訳. 言語処理学会第 30 回年次大会論文集, pp. 2342–2347, 2024.

[3] 加納保昌, 須藤克仁, 中村哲. 分割統治的ニューラル機械翻訳. 言語処理学会第 27 回年次大会論文集, pp. 148–153, 2021.

[4] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **In Proc. EMNLP**, pp. 66–71, 2018.

[5] Y. Liu, et al. Multilingual denoising pre-training for neural machine translation. **TACL**, Vol. 8, pp. 726–742, 2020.

[6] M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In **Proc. LREC-COLING**, pp. 9452–9462, 2024.

[7] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proc. NAACL**, pp. 48–53, 2019.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[9] M. Post. A call for clarity in reporting BLEU scores. In **Proc. 3rd WMT**, pp. 186–191, 2018.

[10] J. Pouget-Abadie, D. Bahdanau, B. van Merrien-Boer, K. Cho, and Y. Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. **In Proc. 8th SSST**, pp. 78–85, 2014.

[11] K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Nagata. Divide and translate: Improving long distance reordering in statistical machine translation. In **Proc. 5th WMT**, pp. 418–427, 2010.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In **Proc. 30th NIPS**, pp. 5998–6008, 2017.

[13] R. Wicks and M. Post. A unified approach to sentence segmentation of punctuated text in many languages. In **Proc. 59th ACL**, pp. 3995–4007, 2021.

[14] Q. Wu, M. Nagata, and Y. Tsuruoka. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In **Proc. 61st ACL**, pp. 11084–11099, 2023.

[15] J. Zhang and T. Matsumoto. Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora. **Applied Sciences**, Vol. 9, No. 10, 2036, 2019.

表 3　提案手法による訳抜け・繰り返しの改善例

訳抜け

入力文
前記温度因子及び/又は前記 NTU が、前記熱交換器 (3) の伝熱面積によって、好ましくは前記熱交換器の長さによって提供され、前記熱交換器 (3) が、好ましくは、フィン付きチューブ形状、コイル形状、及び/又はフィン形状であり、前記流路 (2) の円周を少なくとも部分的に取り囲む、 請求項 2 に記載の極低温冷凍システム (1) 。

参照訳文
Cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably of a finned tube shape, coiled shape, and/or fin shape and at least partially surrounds a circumference of the conduit (2).

ベースライン
The cryogenic refrigeration system according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger.

提案手法
The cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably finned tube-shaped, coil-shaped and/or fin-shaped and at least partially surrounds a circumference of the flow channel (2).

繰り返し

入力文
前記流路遮断バルブは、 内部に前記閉鎖部材を収容し、カンチレバー形状からなる少なくとも 1 つの片持ちばりを備え、円筒形に形成されて、前記連通流路の流入口に挿入されるように設置されるボディー; 及び、 一側は前記片持ちばりから突出形成される係止部により支持され、他側は前記閉鎖部材と接触するように設置されるリング部材; を含み、 前記リング部材は設定された温度以上になると、前記閉鎖部材が中心部を通過するように変形されて、前記閉鎖部材を前記連通流路の内部に向けて移動させる、 ことを特徴とする、請求項 17 に記載のバルブアセンブリ。

参照訳文
The valve assembly of claim 17, wherein the flow path blocking valve includes:a body for containing the blocking member therein, and providing at least one cantilever portion formed in a cantilever shape, the body formed in a cylindrical shape and disposed to be inserted into the inlet of the communication flow path; anda ring member having one side supported by a locking portion protruding from the cantilever portion and the other side disposed to contact the blocking member,wherein the ring member moves the blocking member towards the inside of the communication flow path by deforming the blocking member to pass through a central part of the ring member when the internal temperature exceeds the preset temperature.

ベースライン
The valve assembly of claim 17, wherein the flow path shut-off valve comprises:a body configured to accommodate the closing member therein, having at least one cantilever shaped cantilever ...

提案手法
The valve assembly of claim 17, wherein the passage shutoff valve comprises:a body accommodating the closing member therein, having at least one cantilever formed in a cantilever shape, formed in a cylindrical shape, and installed to be inserted into the inlet of the communication passage; anda ring member having one side supported by a locking portion formed to protrude from the cantilever and the other side installed to be in contact with the closing member,wherein the ring member is deformed such that the closing member passes through a central portion and moves the closing member toward the inside of the communication passage when a set temperature or higher is reached.

A　提案手法による訳抜け・繰り返しの改善例

テストデータの中で提案手法による訳抜け・繰り返しの翻訳改善が見られた出力例を表 3 に示す．

B　並び替え・編集モデルの事前学習による影響

本論文では並び替え・編集モデルの事前訓練に日英双方向の対訳データを使用したが，この事前訓練による最終的な並び替え・編集モデルの精度に対して実験を行った．事前訓練に使用する対訳データは JaParaPat の 2016 年から 2020 年のデータとし，日英双方向の対訳データで訓練したモデルの 2 種類を作成した．これらのモデルに対して，同一の並び替え・編集モデル訓練データを用いてファインチューニングを行い，精度比較を行った．2021 年の特許データから特許請求項のみを抽出したテストデータによる BLEU の評価結果を表 4 に示す．この結果から並び替え・編集モデルには日英両言語の理解力が重要であり，事前訓練に日英双方向の対訳データを使用することで，並び替え・編集の精度を向上させることが示唆された．

表 4　BLEU

事前学習に用いたデータ	BLEU
日英方向のみ	55.0
日英双方向	55.6

C　訳抜け・繰り返しに対しての分析

提案手法が訳抜けや繰り返しといった翻訳ミスを抑制できるかを分析するため，ベースラインモデルおよび提案手法による訳出文と参照訳文の文長比を「2 以上」,「2〜0.5」,「0.5 以下」の 3 種類に分類して傾向を観察した．テストセット全体 (238,902 文) での分類結果を表 5 に示す．この表の中で，ベースラインが 2 以上で提案手法が 2-0.5(左列・真ん中行) のような，片方の訳出は訳抜け・繰り返しの可能性があるがもう片方の訳出ではうまく翻訳ができている 4 つの集合 (表中の「真ん中列・上行」,「左列・真ん中行」,「真ん中列・下行」,「右列・真ん中行」) に着目する．その結果，繰り返しを表す「真ん中列・上行」と比較して「左列・真ん中行」の方が頻度が少ないことから，ベースラインモデルの方が繰り返しの発生が少ないことが考えられる．また，訳抜けの場合は「右列・真ん中行」よりも「真ん中列・下行」の方が頻度が少ないことから提案手法の方が訳抜けの発生数が少ないことが示唆された．

表 5　テストセットにおける訳抜け・繰り返し分析（提案手法 vs. ベースライン）

		ベースライン		
		2 以上	2–0.5	0.5 以下
提案手法	2 以上	1,055	376	4
	2–0.5	273	234,489	900
	0.5 以下	2	515	1,217