特許請求項における並列構造の二言語間対応付け

石丸司[1]　宇津呂武仁[2]　永田昌明[3]
[1]筑波大学 理工学群 工学システム学類　[2]筑波大学 システム情報系 知能機能工学域
[3]NTT コミュニケーション科学基礎研究所

概要

本論文では，特許請求項における並列構造の日英二言語間対応付けを対象とし，二つの対応付け手法を提案する．第一に，デコーダモデル (LLM) を用いた並列構造対応付け手法を提案する．本手法は，対訳文の単語対応から構造を推定する従来手法とは異なり，少数の並列構造対応データで学習した LLM を用い，並列構造とその対応関係を一段階で直接予測する点に特徴がある．第二に，mBERT などの Transformer エンコーダモデルによって得られる単語対応を利用し，ソース文中の並列構造に含まれる単語の対応先が最も多く含まれるターゲット文中の並列構造を対応付け先として決定する，多数決に基づく対応付け手法を紹介する．これらの評価結果により，特許請求項における並列構造の日英方向対応付けでの提案手法の有効性が示された．

1　はじめに

並列構造は多くの文章に頻繁に現れる基本的な文構造の一つである．とりわけ特許請求項は，一文が非常に長く，文中に多数の並列構造を含む文書であることが知られている．現在，日本において国際特許出願を行う際には，特許文書を英語へ翻訳する必要が生じる場合が多いが，このような文構造上の特徴により，機械翻訳によって正確な翻訳を得ることは容易ではない．

本研究の最終的な目的は，付録 A に示すように，特許請求項の原文と翻訳文における並列構造の整合性に着目し，翻訳誤りに関する情報を検出・活用することで，翻訳後編集を通じたより正確な特許請求項翻訳の実現を目指すことである．本論文では，その基盤技術として，日本語特許請求項およびその英語対訳文から並列構造を抽出し，二言語間で対応付ける手法を提案する．

まず，並列構造の抽出手法として，デコーダモデル (LLM) を用いた Translation between Augmented Natural Languages(TANL) [9] 形式による単一言語内の構造予測タスクを提案する．TANL は，入力文に対して，構文構造やラベル境界などの構造情報をタグとして埋め込み，モデルにはタグ付き文の生成を直接学習させることで，単一の生成モデルで多様な構造予測タスクを扱うことを可能にする．本手法により，並列構造の抽出を文生成タスクとして定式化し，CoRec [12] などの Transformer エンコーダモデルに基づく従来手法を上回る抽出精度を達成することを示した．

次に，並列構造の対応付け手法として，二つのアプローチを検討する．第一に，LLM を用いた並列構造対応付け手法を提案する．本手法は，対訳文間の単語対応から構造を推定する従来手法である SpanAlign [7] [1] とは異なり，少数の教師データで学習した LLM を用いて，並列構造とその対応関係を一段階で直接予測する点に特徴がある．第二に，mBERT [10] などの Transformer エンコーダモデルによって得られる単語対応を利用し，ソース文中の並列構造に含まれる単語の対応先が最も多く含まれるターゲット文中の並列構造を対応付け先として決定する，多数決に基づく並列構造対応付け手法を紹介する．これらの評価結果により，LLM による対応付け手法が，単語対応を用いた対応付け手法と比較して，F1 および AER で有効性を示した．

2　関連研究

2.1　並列構造抽出

黒橋ら [14] は，動的計画法を用いた先駆的な手法として，文節間の意味的・表層的な類似度に基づき，長い文の解析において困難な並列範囲の特定を，バランスの取れた文節列の発見として定式化した．

さらに，Teranishi ら [11] は，タスクを等位接続詞の発見，内側境界の同定，外側境界の選択という 3 つの副課題へ分解した分解局所モデルを提案し，CKY アルゴリズムによる推論を導入することで，一

貫性のある大域的な等位構造の最適化を実現し，現代的なニューラルネットワーク手法によって高精度な抽出を可能にしている．

Wang ら [12] は，BERT [2] を用いたエンコーダモデルに基づく並列構造抽出手法を提案している．同手法は，等位接続詞の検出と並列句の境界検出という二つのステップから構成され，最先端手法として知られている．

2.2 構文構造の二言語間対応付け

Nagata ら [7] および Chousa ら [1] は，BERT 型エンコーダモデルを用いたスパン対応付け手法を提案している．同手法は，ソース文中のスパンを Question，ターゲット文書を Context，ターゲット文書中の対応スパンを Answer とみなすことで，文アラインメント問題を SQuAD 型のような質問応答タスクとして定式化するものである．これにより，文境界の不一致や非単調な対応関係を許容したスパン対応付けを実現している．

また，Miao ら [6] は，デコーダモデル (LLM) を用いた単語対応付け手法を提案している．同手法は，LLM を半教師あり学習により訓練することで，高品質な単語アライメントを獲得する手法であり，明示的な並列アノテーションへの依存を低減しつつ，高い対応精度を実現している．これらに対し，本研究では，Miao らの手法を基盤として，デコーダモデル (LLM) を用いてソース文とターゲット文のスパン対応を一段階で推定する手法を提案する．加えて，単語対応情報を利用し，ソース文中のスパンに含まれる各単語の対応先を集約することで，それらが最も多く含まれるターゲット文中のスパンを対応付け先として決定する手法も併せて提案する．これにより，構文的に対応するスパンを明示的に同定しつつ，モデル構成および推論過程の簡素化を図る．

3 並列構造抽出

3.1 データセット

本論文では，並列構造の抽出手法として，TANL 形式に基づき並列構造を明示的に生成するようにデコーダモデル (LLM) を訓練・推論する手法を提案する．本手法の概要を図 1 に示す．

訓練データとして，日本語については Kainoki Treebank [4]，英語については Penn Treebank-3 [5] を用いた．各言語において，並列構造を含む文を 10,658 文，並列構造を含まない文を 4,480 文抽出し，並列構造を含む文については，図 1 に示すように，並列句の前後に特殊タグを挿入することで教師データを作成した．タグに関する詳細情報は付録 B に示す．また，訓練および推論のプロンプトは付録 C に示す．

さらに，評価データとして JaParaPat[8] に収録されている 2020 年版の日英対訳特許請求項から 100 対を抽出し，並列構造抽出性能の評価に用いた．

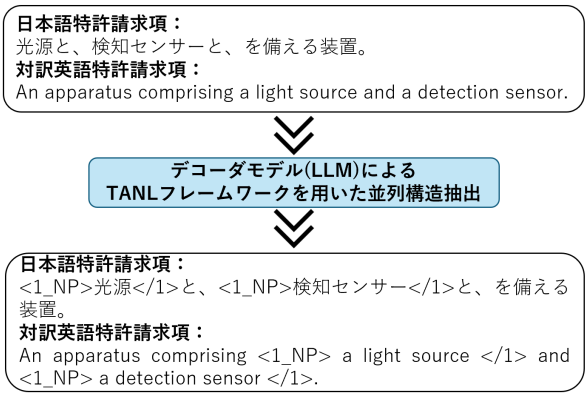

図 1 LLM による TANL フレームワークを用いた並列構造抽出手法

3.2 LLM による並列構造抽出

日本語の並列構造抽出におけるベースモデルとして Qwen3-14B[1] を，英語の並列構造抽出におけるベースモデルとして gemma-3-4b-it[2] を採用した．予備実験として異なるパラメータ数を有する複数のモデルを比較した結果，最も高い精度を示したのは上記 2 つのモデルであった．各モデルに対して，以下の 3 種類の訓練方法を比較した．

- **継続事前訓練 (CPT)**: 7, 569 文を用いた継続事前学習
- **教師ありファインチューニング (SFT)**: 7, 569 文を用いた教師あり微調整
- **CPT→SFT**: それぞれ異なる 7, 569 文を用いて CPT の後に SFT を実施

これらの訓練には，再現性や GPU メモリ容量の観点から，QLoRA を適用している．また，各言語における並列構造抽出性能を比較するため，日本語および英語の両タスクにおいて GPT-5[3]（3-shot）を，さらに英語の並列構造抽出においては従来手法であ

1) https://huggingface.co/Qwen/Qwen3-14B/tree/main
2) https://huggingface.co/google/gemma-3-4b-it
3) https://openai.com/ja-JP/index/introducing-gpt-5/

る CoRec[4)] [12] を比較手法として設定した．

3.3 評価結果

表 1 並列構造抽出における F1 スコア

言語	モデル	F1↑
日	GPT-5(3-shot)	0.7568
	Qwen3-14B(CPT)	0.6732
	Qwen3-14B(SFT)	0.7640
	Qwen3-14B(CPT→SFT)	**0.8481**
英	GPT-5(3-shot)	0.6991
	CoRec	0.7312
	gemma-3-4b-it(CPT)	0.7634
	gemma-3-4b-it(SFT)	**0.8624**
	gemma-3-4b-it(CPT→SFT)	0.7197

評価結果を表 1 に示す．評価指標には F1 スコアを用い，タグを除去した文における並列句単位で算出した．

評価結果より，日本語においては，提案手法である Qwen3-14B を用いた学習済みモデルが，SFT，CPT→SFT の学習設定において GPT-5(3-shot) を上回る性能を示した．特に，CPT に続いて SFT を行う二段階学習 (CPT→SFT) を適用したモデルが F1=0.8481 と最も高い性能を達成しており，CPT と SFT を組み合わせることの有効性が確認できる．

英語においても，提案手法は既存の最先端手法である CoRec を上回る，あるいは同等の性能を示した．gemma-3-4b-it を用いた SFT モデルは F1=0.8624 を達成し，GPT-5(3-shot) および CoRec を明確に上回っている．一方で，CPT→SFT の二段階学習は，英語においては必ずしも最良の結果をもたらさず，単純な SFT が最も高い性能を示した．

これらの結果から，TANL 形式に基づいて並列構造を明示的に生成する本手法は，日本語・英語のいずれにおいても有効であり，教師あり学習によって高精度な並列構造抽出が可能であることが示された．また，言語やモデル規模に応じて，最適な学習戦略 (CPT，SFT，CPT→SFT) が異なる可能性があることも示唆している．

4 並列構造二言語間対応付け

4.1 データセット

本論文では，日英方向への並列構造対応付け手法の一つのアプローチとして，少量の教師データによる LLM の教師あり学習を通じて，並列構造の対応付けを行う手法を提案している．

本手法の訓練データとして，JaParaPat [8] に収録されている 2020 年版の日英対訳特許請求項から 64 対を抽出して用いた．その内訳は，対応付け先を有するデータが 32 対，対応付け先を有しないデータが 32 対である．対応付け先を有するデータについては，(1) ターゲット文中の対応付け対象となるスパンの前後に特殊タグを挿入，または，(2)Juraska ら [3] の提案する最小限のコンテキスト出力方針に基づき，本研究では対応スパンの前後一単語を出力する形式で教師データを作成した．訓練および推論のプロンプトを付録 C に示す．

また，評価データとしては，3.1 節に示した JaParaPat に収録されている 2020 年の日英対訳特許請求項 100 対を用い，並列構造対応付け性能の評価を行った．

4.2 LLM を用いた対応付け

本論文では，並列構造対応付け手法の第一のアプローチとして，Miao ら [6] によって提案された LLM を用いた単語対応付け手法を基盤とし，少量の訓練データによる LLM の教師あり学習を通じて並列構造の対応付けを行う手法を提案する．本手法の概要を図 2 に示す．

対応付けに用いるモデルとして，llama-3-youko-8b[5)]，gemma-3-4b-it[6)]，および Meta-Llama-3.1-8B-Instruct[7)] をベースモデルとし，64 対のデータセットを用いて教師ありファインチューニング (SFT) を行った．これらの SFT は Miao らの手順に従い，LoRA を適用して実施している．

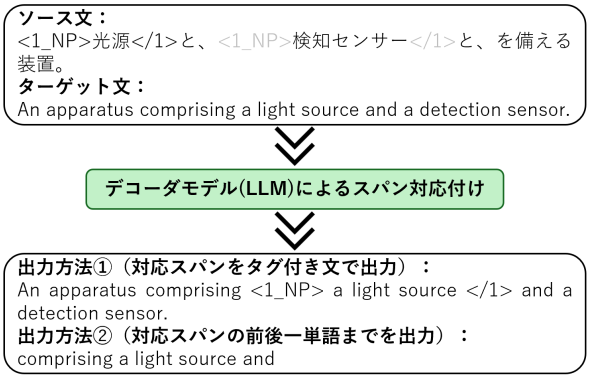

図 2 LLM によるスパン対応付け手法

4) https://github.com/qingwang-isu/CoRec?tab=readme-ov-file

5) https://huggingface.co/rinna/llama-3-youko-8b

6) https://huggingface.co/google/gemma-3-4b-it

7) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

4.3 単語対応を用いた対応付け

本論文では，並列構造対応付け手法の第二のアプローチとして，mBERT [10] などの Transformer エンコーダモデルに基づいて推定される単語対応情報を活用し，多数決に基づく並列構造対応付け手法を提案する．本手法では，ソース文中の並列構造に含まれる各単語の対応先を集約し，それらが最も多く含まれるターゲット文中の並列構造を，当該ソース並列構造の対応付け先として決定する．本手法の概要を図 3 に示す．

具体的に，単語対応付けには，Wu らによって提案された WSPAlign [13] を用いる．並列構造抽出モデルによってあらかじめタグ付けされたソース文およびターゲット文に対して単語対応を推定し，その結果をスパンレベルの対応付けに利用する．Wu らの手順に従い，日英方向の単語対応付けモデルとして，mBERT をベースとする WSPAlign-mbert-base[8] および WSPAlign-ft-kftt[9] の二種類を用いた．

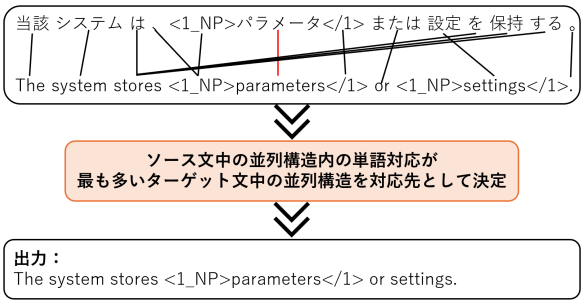

図 3 単語対応を用いた多数決によるスパン対応付け手法

4.4 評価結果

表 2 並列構造対応付けにおける F1 スコアと AER．(1)LLM による対応スパンをタグ付き文で出力．(2)LLM による対応スパンの前後一単語までを出力．(3) 単語対応を用いた対応付け．∗ は抽出から対応付けを一貫して行った場合のスコアであり，抽出は表 1 の最高精度モデルを用いた．

手法	モデル	F1↑	AER↓	F1∗↑	AER∗↓
(1)	llama-3-youko-8b	**0.6800**	**0.3199**	**0.6253**	**0.3747**
	gemma-3-4b-it	0.3936	0.6046	0.3089	0.6911
	Meta-Llama-3.1-8B-Instruct	0.5078	0.4894	0.4530	0.5799
(2)	llama-3-youko-8b	0.3655	0.6165	0.3422	0.6320
	gemma-3-4b-it	0.1074	0.8391	0.0923	0.9077
	Meta-Llama-3.1-8B-Instruct	0.4237	0.5788	0.3683	0.6317
(3)	WSPAlign-mbert-base	0.4938	0.4986	0.4070	0.5770
	WSPAlign-ft-kftt	0.4800	0.5126	0.4230	0.5930

8) https://huggingface.co/qiyuw/WSPAlign-mbert-base
9) https://huggingface.co/qiyuw/WSPAlign-ft-kftt

評価結果を表 2 に示す．評価指標には，F1 スコアと Alignment Error Rate(AER) を用い，タグを除去した文における並列句単位で算出した．

評価結果より，LLM を用いた対応付け手法において，対応スパンをタグ付き文として出力する出力形式 (1) は，出力形式 (2) や単語対応に基づく手法と比較して，高い F1 スコアおよび低い AER を達成する傾向が確認された．特に，llama-3-youko-8b を用いたモデルは最も高い性能を示した．

一方，単語対応に基づく対応付け手法では，特許請求項が一文あたり非常に長いという文書特性により，単語対応付けの精度が低下し，並列構造対応付けが困難になる可能性が示唆される．また，タグ付けされた文において単語対応を行ったことによる精度低下が考えられる．加えて，LLM を用いた対応付け手法においても，出力形式の違いによって性能差が生じた．

さらに，並列構造抽出から対応付けまでを一貫して行った場合，対応付け精度が抽出精度に強く依存するため，すべての手法において性能低下が観測された．このことから，実用化に向けては，抽出精度のさらなる向上が重要な課題である．

以上より，少量の教師データを用いた LLM の教師あり学習に基づく一段階の並列構造対応付け手法は，単語対応情報に依存する手法と比較して，高精度かつ安定した対応付けを実現できることが示された．一方で，抽出から対応付けまでを一貫して運用する際の課題も明らかとなった．

5 おわりに

本論文では，特許請求項における並列構造の日英二言語間対応付けを対象とし，並列構造の抽出および対応付けを扱う手法を提案した．まず，TANL 形式に基づき，LLM を用いて並列構造を明示的に生成する並列構造抽出手法を提案し，並列構造抽出を一段階の生成問題として定式化する本手法の有効性を示した．

次に，並列構造の二言語間対応付けに関して，(i)LLM を教師あり学習し，並列構造の対応関係を一段階で推定する手法と，(ii) 単語対応を用いた多数決に基づく対応付け手法の二つのアプローチを検討した．評価実験の結果，LLM を用いた対応付け手法は，単語対応に基づく手法と比較して，高い F1 スコアおよび低い AER を達成し，並列構造の二言語間対応付けにおいて本手法の有効性を示した．

参考文献

[1] K. Chousa, M. Nagata, and M. Nishino. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In **Proc. 28th COLING**, pp. 4750–4761, 2020.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. NAACL**, pp. 4171–4186, 2019.

[3] J. Juraska, T. Domhan, M. Finkelstein, T. Nakagawa, G. Kovacs, D. Deutsch, P. Wang, and M. Freitag. MetricX-25 and GemSpanEval: Google translate submissions to the WMT25 evaluation shared task. In **Proc. WMT**, pp. 957–968, 2025.

[4] E. Kainoki. The Kainoki treebank. https://kainoki.github.io, 2022.

[5] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Penn treebank-3, 1999.

[6] Z. Miao, Q. Wu, M. Nagata, and Y. Tsuruoka. Improving word alignment using semi-supervised learning. In **Findings of ACL**, pp. 19871–19888, 2025.

[7] M. Nagata, K. Chousa, and M. Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In **Proc. EMNLP**, pp. 555–565, 2020.

[8] 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.

[9] G. Paolini, B. Athiwaratkun, J. Krone, J. Ma, A. Achille, R. Anubhai, C. Nogueira dos Santos, B. Xiang, and S. Soatto. Structured prediction as translation between augmented natural languages. In **Proc. ICLR**, pp. 1–26, 2021.

[10] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? **arXiv:1906.01502**, 2019.

[11] H. Teranishi, H. Shindo, and Y. Matsumoto. Decomposed local models for coordinate structure parsing. In **Proc. NAACL-HLT**, pp. 3394–3403, 2019.

[12] Q. Wang, H. Jia, W. Song, and Q. Li. CoRec: An easy approach for coordination recognition. In **Proc. EMNLP**, pp. 15112–15120, 2023.

[13] 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.

[14] 黒橋禎夫, 長尾真. 長い日本語文における並列構造の推定. 情報処理学会論文誌, Vol. 33, No. 8, pp. 1022–1031, 1992.

A　本研究における最終目的および全体像

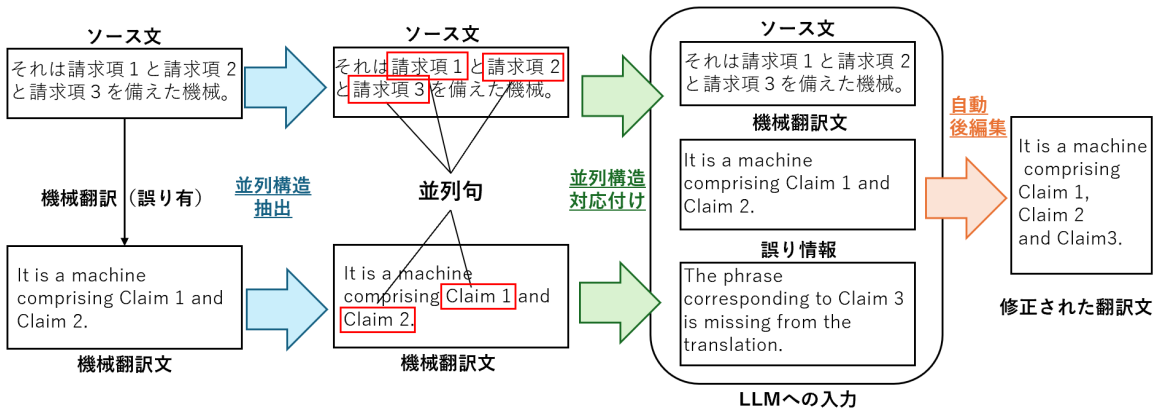

図 4　特許請求項における日英二言語間の並列構造整合性に着目した自動翻訳後編集

B　並列構造タグのアノテーション

本論文では，並列構造を表すタグとして，並列句を<∗_XP>および</∗>で囲む形式のアノテーションを採用する．ここで，∗ は文中に存在する各並列構造に一意に割り当てられる番号を表し，Kainoki-Treebank および PennTreebank-3 における構文木の浅い順に基づいて付与される．また，XP は並列構造を構成する句の品詞ラベルを示す．なお，日本語と英語では，並列句として認識される句ラベルの種類が一部異なる．

C　プロンプト例

並列構造抽出モデルのプロンプト例

Instruction: 以下の文に含まれる並列構造をタグ付きで抽出してください。
Input: 光源と、検出センサーと、を備える証明装置。
Output:
“‘ <1_NP>光源</1>と、<1_NP>検出センサー</1>と、を備える証明装置。”’

並列構造対応付けモデルのプロンプト例 (括弧内は手法 (2))

Role: system
Content: You are a helpful AI assistant for span alignment.
Role: user
Content: Please indicate the corresponding portion in the English text enclosed by <∗_XP> and </∗> for the span enclosed by <∗_XP> and </∗> in the following Japanese text. If there is no corresponding span, do not indicate it. (Please indicate the corresponding span and the word im。mediately before and after it for the span enclosed by <∗_XP> and </∗> in the following Japanese text. If there is no corresponding span, do not output anything.)
The parallel sentences:
<1_NP>光源</1>と、検出センサーと、を備える証明装置。
A lighting apparatus comprising a light source and a detection sensor.
Role: assistant
Content: Here is the span alignment information:
“‘A lighting apparatus comprising <1_NP> a light source </1> and a detection sensor. (comprising a light source and)”’