逐次文分類の言語間転移では言語学的な近さよりも修辞構造の類似性が寄与する

山内一礼 [1]　桂井麻里衣 [1]
[1] 同志社大学大学院 理工学研究科
{yamauchi23, katsurai}@mm.doshisha.ac.jp

概要

逐次文分類（SSC）は，学術論文を構造化する上で重要なタスクである．しかし先行研究は英語のみを対象にしており，より広範な科学的知見へのアクセスを可能にするには，他言語への拡張が求められる．従来の多言語処理研究では，言語学的な近さが言語間転移を成功させることが示されてきた．一方本研究では，SSC において修辞構造の類似性が言語間転移に強く寄与するという仮説を立てる．これを検証するため，我々は多言語 SSC データセットを構築し分析を行った．その結果，構造的類似性は言語学的近接性より転移性能と強く相関することが示された．この知見に基づき構造情報を活用する手法を提案し，言語横断的な SSC で性能向上を達成した．

1　はじめに

逐次文分類 (sequential sentence classification; SSC) は，学術論文の各文を背景・目的・手法・結果・結論などの修辞的役割に分類するタスクであり，文献検索や自動要約などの下流タスクを支える基盤技術である．中でも学術論文のアブストラクトを対象とした SSC は盛んに研究されており，Transformer ベースのモデル [1] や大規模言語モデル (LLM) [2] の発展により性能が向上している．しかしながら，先行研究の大半は英語を対象としており，非英語論文への SSC 適用は依然として課題である．

多言語処理の研究では，多言語事前学習モデルを用いた言語間転移学習が有望なアプローチとして注目されている．これらの研究では，言語間転移の成功は言語学的な近接性に依存することが示されてきた [3, 4]．しかし近年，その有効性はタスクによって異なることが報告されており [5]，SSC においても言語学的近接性以外の要因が言語間転移に寄与する可能性がある．

本研究では，SSC 特有の性質に着目する．学術論文のアブストラクトは，背景が冒頭，結論が末尾に配置されるなど，異なる言語間で類似した論理的進行パターンを持つことが知られている [6, 7]．我々は，このような**構造的類似性が言語間転移に寄与する**という仮説を立て，多言語 SSC データセットを構築して検証する．実験の結果，構造的類似性は従来の言語学的近接性より転移性能と強く相関することが明らかになった．この知見に基づき，我々は構造情報を明示的に活用する三つの手法を提案する．提案手法は macro-F1 指標において既存ベースラインからの性能向上を達成し，言語横断的 SSC における構造情報の有効性を示した．

2　多言語 SSC データセット

本研究では，文献 [8] のアプローチに従い，アブストラクト中に明示的に記載されたセクションヘッダ（例：Background, Method, Result）を正解ラベルとして採用することで，大規模な多言語データセットを構築した．データソースとして，API 経由でのデータ取得が許可されている学術データベース（DOAJ[1)]，HAL[2)]，Dialnet[3)]，TRdizin[4)]，CiNii Research[5)]）を使用し，13 の非英語言語を対象とした．各言語において，“Method,” “Result” を翻訳したものを検索クエリとし，正規表現によりセクションヘッダの存在を確認した．前処理として，HTML エンティティ変換，Unicode 正規化，langdetect [9] ライブラリによる言語検出，重複除去を行い，文分割には NLTK [10]（欧州言語）および spaCy [11]（アジア言語）を使用した．また，2 セクション以上を含むアブストラクトのみを保持した．最終的に，英語の既存データセットである PubMed-RCT 20k [8] を含む

1) https://doaj.org/
2) https://hal.science/
3) https://dialnet.unirioja.es/
4) https://trdizin.gov.tr/
5) https://cir.nii.ac.jp/

表 1　データセット統計

言語	ソース	論文数	文数
英語	PubMed-RCT	20,000	180,040
フランス語	HAL	11,210	134,393
日本語	CiNii	8,366	78,843
スペイン語	Dialnet	5,768	55,743
中国語	DOAJ	3,522	24,649
ロシア語	DOAJ	1,163	9,522
ポルトガル語	Dialnet	1,122	8,865
イタリア語	DOAJ	624	6,353
インドネシア語	DOAJ	434	4,270
トルコ語	TRdizin	179	630
韓国語	DOAJ	48	485
ポーランド語	DOAJ	30	369
オランダ語	DOAJ	14	131
エストニア語	DOAJ	7	123

14 言語で 47,487 件のアブストラクト，504,416 文からなるデータセットを構築した．各言語の詳細な統計を表 1 に示す．

3　言語間転移の分析

3.1　実験設定

言語間転移に寄与する要因を分析するため，zero-shot 言語間転移実験を行った．zero-shot 言語間転移とは，ソース言語のみで訓練したモデルを，訓練時に使用していないターゲット言語のテストデータで評価することを指す．モデルとして，多言語 BERT（mBERT） [12] を基盤とし，hierarchical sequential labeling network（HSLN） [13] による階層的系列ラベリングを行う mBERT-HSLN [1] と，LLM の Qwen2.5-3B-Instruct [14] を使用した．LLM のプロンプトは，プロンプト言語の影響を排除するため英語で統一した（付録 B）．実験には，200 件以上のアブストラクトを持つ 9 言語（中国語，スペイン語，英語，フランス語，インドネシア語，イタリア語，日本語，ポルトガル語，ロシア語）を用い，これら 9 言語のすべてのペア（同一言語を含む）81 組で実験した．各言語のデータは訓練 70%，検証 15%，テスト 15% に分割した．テストセットが 200 件を超える言語については，計算コストと評価の信頼性のバランスを考慮し，ランダムに 200 件をサンプリングした．各言語ペアについてランダムにシードを変えて 3 回実験を行い，平均 macro-F1 を算出した．

3.2　類似度指標の定義

言語間の類似度指標として言語学的近接性と構造的類似性の 2 種類を定義した．まず**言語学的近接性**として，lang2vec [15] によって提供されている統語，音韻，目録，地理，系統の 5 カテゴリの特徴ベクトルを連結し，言語ペア間のコサイン類似度を計算した．この指標では，イタリア語とポルトガル語のような類型論的に近い言語ペアで高い値（0.932）を示し，日本語とポルトガル語のような遠い言語ペアで低い値（0.647）を示した．

構造的類似性として，アブストラクトの修辞構造の類似性を測るため，(1) 5 ラベルの出現頻度分布の Jensen-Shannon divergence（JSD），(2) ラベルごとのセクション長（同一ラベルの連続文数）分布の JSD，(3) ラベル継続確率（次の文が同じラベルを持つ確率）のユークリッド距離，(4) 主要なセクション遷移（例：手法→結果）の平均相対位置の差異，(5) ラベルごとのブロック数（非連続スパンの数）分布の JSD，(6) ラベル遷移分布のシャノンエントロピーの差異．の 6 種類の指標を定義した（詳細は付録 A）．これらの平均を全体の構造的距離 d とし，構造的類似性を $1 - d$ と定義した．

3.3　相関分析の結果

図 1 に，Qwen2.5-3B の転移性能行列と 2 種類の類似度行列を示す．図 2 に転移性能と 2 種類の類似度の散布図と回帰直線を示す．同一言語ペア（9 ペア）を除く 72 ペアでピアソン相関係数を算出したところ，構造的類似性と転移性能の間には両モデルで統計的に有意な正の相関が見られた（mBERT-HSLN: $r = 0.408$, $p < 0.001$; Qwen2.5-3B: $r = 0.343$, $p < 0.001$）．一方，言語学的近接性との相関は両モデルで有意でなかった（mBERT-HSLN: $r = 0.087$, $p = 0.438$; Qwen2.5-3B: $r = -0.042$, $p = 0.726$）．この結果は，SSC においてモデルが言語学的な特徴よりも，「目的は背景の後に来る」「結論は結果の後に来る」といった修辞構造に依存していることを示唆する．実際，日本語とスペイン語・ポルトガル語は全く異なる言語族に属するにもかかわらず，構造的類似性が高く（0.93），実際の転移性能も日本語からスペイン語は 0.602，日本語からポルトガル語は 0.632 と良好であった．

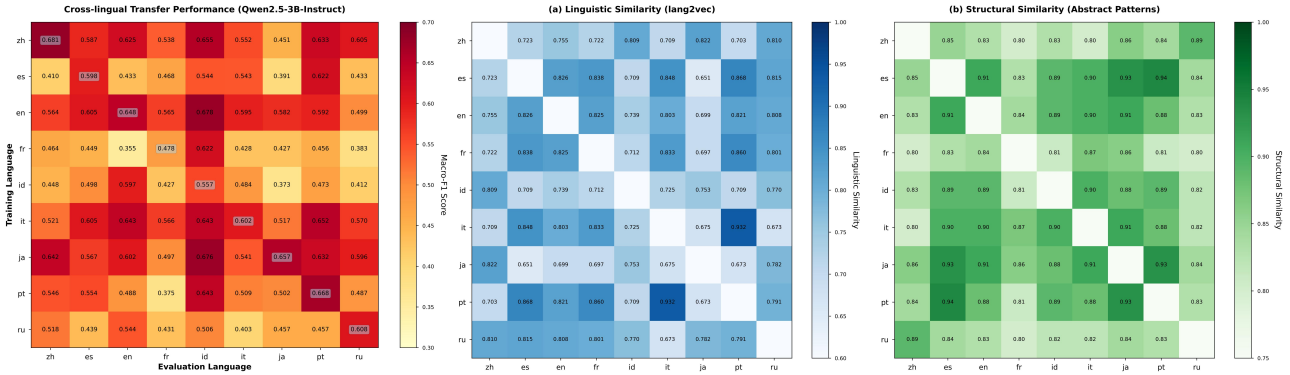

図 1　言語間転移の分析結果．左：Qwen2.5-3B の転移性能行列（macro-F1）．中央：言語学的近接性の行列．右：構造的類似性の行列．

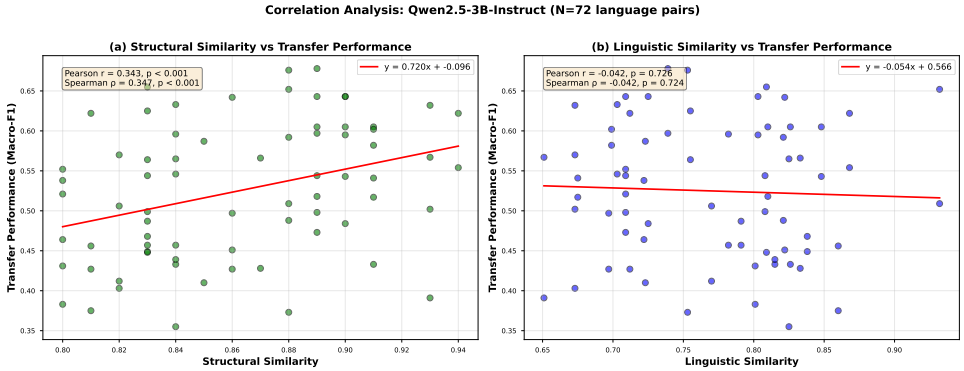

図 2　2 種類の類似度と Qwen2.5-3B における転移性能の相関（$N = 72$）．左：言語学的近接性と転移性能の散布図（$r = -0.042$, $p = 0.726$）．右：構造的類似性と転移性能の散布図（$r = 0.343$, $p < 0.001$）．構造的類似性は有意な正の相関を示す一方，言語学的近接性は有意な相関を示さない．

4　構造情報の活用

前節の分析結果に基づき，構造情報を明示的に活用して SSC の性能を向上させる三つの手法を提案する．

4.1　Structure-informed prompting (SIP)

SIP は，アブストラクトの典型的な構造を LLM のプロンプトに明示的に組み込む手法である．具体的には，タスク説明とラベル定義に加えて，「学術論文のアブストラクトは通常，背景，目的，手法，結果，結論の順序で構成される」という構造的制約をプロンプトに含める（付録 B）．これにより，モデルは個々の文の内容だけでなく，アブストラクトの構造を考慮した予測が可能となる．

4.2　Structure-guided verifier reranking (SGVR)

SGVR は，LLM が生成した複数の候補予測を，構造的特徴に基づいて学習した検証器でリランキングする手法である．機械翻訳における品質推定のアプローチ [16] を参考に，推論時に正解ラベルを用いずに出力品質を予測する検証器を訓練する．具体的には，まず温度サンプリング（温度=0.7）により訓練データの各アブストラクトに対して K 個の候補ラベル系列を生成する．次に，各候補から 44 次元の構造的特徴ベクトルを抽出する．特徴量は，ラベル分布特徴，遷移特徴，位置特徴から構成される（詳細は付録 C）．検証器には LightGBM [17] を採用し，これらの構造的特徴を入力として候補の文レベル正解率を予測する回帰モデルを訓練する．訓練時には，各候補について正解ラベルと比較して計算した文レベルの正解率を教師信号として用いる．推論時には，検証器が最も高い正解率を予測した候補を最終予測として選択する．

4.3　Structure-adaptive verifier (SAV)

SAV は，ターゲット言語の訓練データが利用できない zero-shot 設定での教師なし手法である．Self-consistency [18] の考え方を拡張し，複数回のサンプリングで一貫して予測されるラベルはより正確である可能性が高いという仮定に基づく．温度サンプリングにより各アブストラクトに対して K 個の

表 2　実験結果．訓練言語での評価（左）と zero-shot 評価（右）．

訓練言語での評価			zero-shot 評価		
手法	正解率	Macro-F1	手法	正解率	Macro-F1
mBERT-HSLN	0.919	0.812	mBERT-HSLN	0.751	0.658
SIP+SGVR (Qwen)	0.923	0.848	SIP+SAV (Qwen)	0.849	0.818

候補ラベル系列を生成し，アブストラクト内の各文の位置（以下，「位置」と呼ぶ）に着目して各候補を三つのスコアで評価する．

1. 一貫性スコア：評価対象の候補と他の候補との位置ごとの一致率の平均である．
2. 信頼度スコア：各位置において，最頻ラベルと一致する場合はその出現率，異なる場合は予測ラベルの出現率を用いて計算し，全位置で平均したものである．
3. 遷移スコア：個別のラベルではなく隣接ラベル間の遷移パターンに対して信頼度スコアと同様の計算を行ったものである．

これら 3 スコアの平均が最も高い候補を最終予測として選択する（詳細は付録 D）．位置ごとに最頻ラベルを選ぶ単純な多数決とは異なり，SAV は候補集合内に実際に存在する系列のみを出力できる．

5　提案手法の有効性検証実験

5.1　設定

本章では 3 章で使用した 9 言語の混合データで訓練したモデルを評価する．評価設定は 2 種類で，(1) 訓練言語での評価と (2) zero-shot 評価（未訓練言語でのテスト）である．訓練言語での評価のため，9 言語それぞれを 70 : 15 : 15 で訓練・検証・テスト用に分割し，全言語の訓練データを統合してモデルを訓練した．zero-shot 評価では，訓練に使用していない 5 言語（エストニア語，韓国語，オランダ語，ポーランド語，トルコ語）で評価した．Qwen2.5-3B-Instruct に LoRA [19]（$r = 32, \alpha = 64$）を適用し，学習率 2×10^{-4}，バッチサイズ 4，3 エポック，プロンプトは SIP を用いて訓練した．ハイパーパラメータは検証セットでの性能に基づき選択した．訓練言語での評価における SGVR では $K = 3$，zero-shot 評価における SAV では候補の多様性と計算量の兼ね合いから $K = 5$ とした．ベースラインとして mBERT-HSLN を使用し，そのハイパーパラメータは文献 [1] に倣った．評価指標は文レベルの正解率と macro-F1 である．

5.2　結果

表 2 に実験結果を示す．訓練言語での評価において，SIP+SGVR（Qwen）は正解率 0.923，macro-F1 0.848 を達成し，mBERT-HSLN（正解率 0.919，macro-F1 0.812）に対して Macro-F1 で +0.036 の改善を示した．言語別詳細結果（付録 E）を見ると，インドネシア語は正解率 0.960，macro-F1 0.955 であり非常に高い精度で分類できていることがわかる．逆に，ロシア語や中国語はそれぞれ macro-F1 が 0.751 と 0.788 で低かった．zero-shot 評価では，SIP+SAV（Qwen）が mBERT-HSLN に対して正解率で +0.098，macro-F1 で +0.016 の改善を達成した．言語別詳細結果（付録 E）を見るとポーランド語が正解率 0.894，macro-F1 0.887 と高く，トルコ語が正解率 0.705，macro-F1 0.608 と低かった．これらの結果から，プロンプトでの明示的な構造情報の提示や構造的特徴量を用いた検証器によるリランキングが，多言語 SSC の性能向上に有効であることが示された．

6　まとめ

本研究では，14 言語からなる多言語 SSC データセットを構築し，そのうち 9 言語の全ペア（81 通り）で言語間転移実験を実施して転移性能と各種類似度指標の相関を分析した．その結果，従来重視されてきた言語学的近接性よりも修辞構造の類似性が転移性能と強く相関することが BERT と LLM の両モデルで観察された．この知見に基づき提案した三つの手法（SIP, SGVR, SAV）は，構造情報を明示的に活用することで，訓練言語での評価と zero-shot 評価の両方で性能向上を達成した．提案手法は訓練言語での評価と zero-shot 評価の両設定で性能向上を達成し，明示的な構造情報の活用が多言語 SSC において有効であることを実証した．今後の課題として，他の談話レベルタスクへの手法の拡張を予定している．

謝辞

本研究は JPSP 科研費 JP25K03419 の助成を受けたものです．本研究は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施しました．

参考文献

[1] Arthur Brack, Elias Entrup, Markos Stamatakis, Pascal Buschermöhle, Anett Hoppe, and Ralph Ewerth. Sequential sentence classification in research papers using cross-domain multi-task learning. *International Journal on Digital Libraries*, Vol. 25, pp. 377–400, 2024.

[2] Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16086–16104, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[3] Fred Philippy, Siwen Guo, and Shohreh Haddadan. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5877–5891, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8653–8684, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] Chloé Braud, Maximin Coavoux, and Anders Søgaard. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 292–304, Valencia, Spain, April 2017. Association for Computational Linguistics.

[7] Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, Vol. 54, No. 2, pp. 587–613, 2020.

[8] Franck Dernoncourt and Ji Young Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 308–313, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[9] Nakatani Shuyo. Language detection library for Java, 2010.

[10] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[11] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[13] Di Jin and Peter Szolovits. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3100–3109, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[14] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[15] Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, 2017.

[16] Lucia Specia and Kashif Shah. Machine translation quality estimation: Applications and future perspectives. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pp. 201–235, Cham, 2018. Springer.

[17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 3146–3154. Curran Associates, Inc., 2017.

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

A　構造的類似性指標の詳細

本文で述べた 6 種類の構造的類似性指標について，計算方法の詳細を示す．**(1) ラベル分布:** 各言語 L でラベル l の出現確率 $P_L(l)$ を計算し，$d_{\text{label}}(L_1, L_2) = \text{JSD}(P_{L_1}, P_{L_2})$ とした．**(2) セクション長分布:** 各ラベル l について，セクション長 k の分布 $P_{L,l}(k)$ を求め，$d_{\text{section}} = \frac{1}{5}\sum_l \text{JSD}(P_{L_1,l}, P_{L_2,l})$ とした．**(3) 継続確率:** 各ラベルの継続確率 $p_L(l) = P(y_{i+1} = l | y_i = l)$ を 5 次元ベクトルとし，正規化ユークリッド距離 $d_{\text{cont}} = \frac{1}{\sqrt{5}}\sqrt{\sum_l (p_{L_1}(l) - p_{L_2}(l))^2}$ を計算した．**(4) 境界位置:** 4 つの主要遷移 t（B→O, O→M, M→R, R→C）について，アブストラクト内の平均相対位置 $\mu_L(t)$ を算出し，$d_{\text{bound}} = \frac{\sum_t w_t |\mu_{L_1}(t) - \mu_{L_2}(t)|}{\sum_t w_t}$ とした．ここで $w_t = \min(\text{freq}_{L_1}(t), \text{freq}_{L_2}(t))$ は頻度による重み付けである．**(5) ブロック数分布:** 各ラベルのブロック数（非連続出現回数）分布 $P^{\text{block}}_{L,l}$ を求め，$d_{\text{block}} = \frac{1}{5}\sum_l \text{JSD}(P^{\text{block}}_{L_1,l}, P^{\text{block}}_{L_2,l})$ とした．**(6) 遷移エントロピー:** 全 25 パターンの遷移確率からシャノンエントロピー $H_L = -\sum_{l_1,l_2} P_L(l_2|l_1) \log_2 P_L(l_2|l_1)$ を計算し，$d_{\text{entropy}} = |H_{L_1} - H_{L_2}| / \log_2 25$ とした．

B　プロンプト

LLM 実験で使用したプロンプトを以下に示す．上が 3 章で使ったプロンプト，下が SIP である．

Instruction: 'You must categorize the given sentence into one of these five labels: BACKGROUND, OBJECTIVE, METHOD, RESULT, CONCLUSION. Respond with ONLY the label name.' **Output:** 'Question: What is the rhetorical role of the Target Sentence? Answer with one word from the labels list.'

Instruction: You must categorize the given sentence into one of these five labels: Background, Objective, Method, Result, Conclusion. Respond with ONLY the label name.
Structural Constraints: Academic abstracts typically follow this order: Background (introducing the topic) → Objective (stating the research goal) → Method (describing the approach) → Result (presenting findings) → Conclusion (summarizing implications).
Example: [One demonstration example from training data]
Output: Question: What is the rhetorical role of the Target Sentence? Answer with one word from the labels list.

C　SGVR の構造的特徴量

44 次元の特徴ベクトルは以下から構成される: **1. ラベル分布:** 各ラベルの出現回数 (5)，候補と訓練データのラベル分布のコサイン類似度 (1)，セクション長の平均・標準偏差・最大値 (3)．**2. 遷移:** 隣接ラベル遷移の平均対数確率 (1)，主要遷移（BACKGROUND → OBJECTIVE，OBJECTIVE → METHOD，METHOD → RESULT，RESULT → CONCLUSION）のバイナリ指標 (4)，遷移エントロピー (1)，各ラベルの継続確率 (5)．**3. 位置:** 各ラベルが期待文位置範囲内に出現するかのスコア (5)，開始・終了ラベルの one-hot 表現 (10)．**4. 構造:** 各ラベルの連続ブロック数 (5)，ブロック長の平均・標準偏差 (5)．

D　SAV で使用したスコアの詳細

一貫性スコアは，評価対象の候補と他の各候補との位置ごとの一致率を計算し，$K - 1$ 個の候補について平均したものである．候補が他の候補と平均して何%一致しているかを測定する．

信頼度スコアは，各位置について最頻ラベルとその出現率を計算し，評価対象の候補が最頻ラベルを予測している場合はその出現率，異なるラベルを予測している場合はそのラベルの出現率をスコアとし，全位置で平均する．例えば，10 個の候補のうち 8 個が「背景」を予測する位置で，評価対象も「背景」ならスコアは 0.8,「目的」（2 個のみ）ならスコアは 0.2 となる．

遷移スコアは，信頼度スコアと同じ方法を，個別のラベルではなく隣接ラベル間の遷移パターンに適用したものである．各遷移位置について最頻遷移パターンとその出現率を計算し，信頼度スコアと同様の方法でスコアを算出して平均する．

E　言語別詳細結果

表 3 に訓練言語での評価の言語別結果を示す．表 4 に zero-shot 評価の言語別結果を示す．

表 3　訓練言語での評価の言語別結果（SIP+SGVR, Qwen）

言語	正解率	Macro-F1
英語	0.934	0.875
フランス語	0.943	0.838
日本語	0.866	0.796
スペイン語	0.899	0.882
中国語	0.985	0.788
インドネシア語	0.960	0.955
ポルトガル語	0.881	0.862
イタリア語	0.871	0.840
ロシア語	0.940	0.751
全体	**0.923**	**0.848**

表 4　zero-shot 評価の言語別結果（SIP+SAV, Qwen）

言語	正解率	Macro-F1
エストニア語	0.919	0.709
韓国語	0.981	0.787
オランダ語	0.863	0.838
ポーランド語	0.894	0.887
トルコ語	0.705	0.608
全体	**0.849**	**0.818**