大規模言語モデルを用いた物語分析データセットの効率的構築: 日本語物語の話者推定を例として

郷原聖士　上垣外英剛　渡辺太郎
奈良先端科学技術大学院大学
{gobara.seiji.gt6,kamigaito.h, taro}@is.naist.jp

概要

物語分析におけるセリフの話者推定は，物語の中で登場人物の心情変化や成長を分析していく上で重要なタスクである。しかし，話者推定は複雑な対話，発話パターンの多様性，および曖昧なキャラクター参照のような様々な要素が組み合わさっているため，データセットを人手で作成する場合、大量の作業時間と費用が必要になる．本研究では，大規模言語モデル（LLM）と少量の人手修正を組み合わせたラベル付けをすることで，効率良く高品質なデータセットを構築する手法を示す．実験では，人手でアノテーションした一部の少数の例を LLM に回答例として入力し，最小限の人間による修正を通じてデータセットを構築した．その結果，LLM は設定と話の展開が複雑な「三国志」という物語において約 90%の精度で話者名を正確に識別し，人手のアノテーションコストを抑えられることが示唆された．

1　はじめに

物語分析は，文化的価値観，心理的動態，創造的プロセスを理解する上で不可欠である．物語の構造やテーマを分析することで，社会規範や人間行動に関する知見が得られることが知られている [1]．大規模言語モデル（LLM）は，物語分析においても新たな可能性をもたらしており，登場人物の感情分析やプロット進行予測などの幅広いタスクへの適用可能性を示している [2]．

話者推定は物語分析における重要なタスクであり，台詞を登場人物へ正確に対応付け，物語内でのキャラクター間の動態を理解することが求められる．しかし，高品質な話者推定用データセットを構築するのはコストと労力を要し，一貫性や言い換え表現への注意が必要となる [3, 4, 5, 6, 7]．

また，既存研究では，日本語小説を対象とした話者推定タスクに取り組む例があるものの [8, 9]，使用されているデータセットの多くは非公開か，ライトノベルの特定ジャンルに限定されており，権利上の制約が多い．さらに，小説から話者とセリフを自動で識別する研究も存在するが [10]，表層的な表現の違いに影響を受けやすいという課題がある．

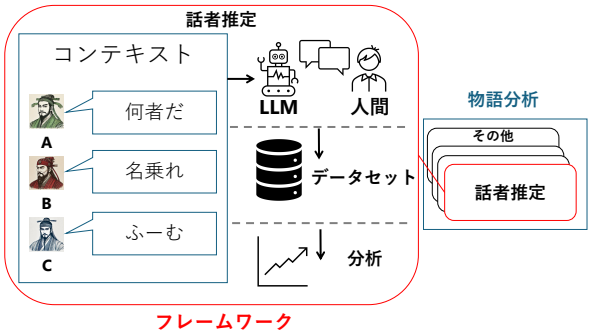

図 1: 話者推定におけるデータセット構築の全体フロー．LLM による初期アノテーションと人手修正を組み合わせ，物語分析のための高品質なデータセットを効率的に構築する手法のフローを示す．

そこで本研究では，LLM によるデータセット構築手法 [11] を物語分析に対しても適用し，LLM による初期アノテーションと人手修正を組み合わせることで，低コストで高品質なデータセットを構築した．図 1 に，LLM による予測と人間による修正を反復的に組み合わせる本研究の処理の流れを示す．データセット構築は，（1）台詞とコンテキストの抽出，（2）LLM による話者推定，および（3）人手による精査という 3 段階で構成されている．

実験の結果，LLM を用いた日本語小説の話者推定のためのデータセット構築では，90%の高精度で話者名を識別できることが明らかになった．また，LLM によって識別された話者名を元にしてデータセットを修正することで，全てのデータセットを人手で構築する場合に比べて，アノテーションにかかる人間の作業時間を減少させられることが示唆された．

2 関連研究

2.1 データセット構築

Elson ら [3] は，19 世紀の英語物語 11 作品において話者名と性別をアノテーションした．He ら [4] は，書籍「*Pride & Prejudice*」において区切られた行を 1 つの発話として扱い，アノテーションを行った．Muzny ら [5] は，これらのデータセットを拡張し，3 つの物語における全ての発話に注釈を付与した QuoteLi3 データセットを構築した．Chen ら [6] は，中国語の物語 World of Plainness (WP) に対する話者推定用データセットを構築した．Vishnubhotla ら [7] は，Project Dialogism Novel Corpus (PDNC) を開発し，28 の英語小説にわたり，話者名とその言い換え表現，対話相手，引用タイプ，参照表現，およびメンションの情報などの，多様な観点からの注釈を付与した．既存のデータセットは主に英語または中国語に限定されており，日本語の公開データセットは僅かである．さらに，これらのデータセットは手動でのアノテーションに依存しているため，本質的に労働集約的かつ高コストである．

2.2 話者推定

特徴量ベースの手法　先行研究 [3, 4, 5, 12] では，話者推定のために言語的特徴や手作業による属性設定を用いる手法を提案している．

機械学習ベースの手法　深層学習の台頭に伴い，話者推定を機械学習ベースで実施する手法が提案されている．例えば，事前学習済み言語モデルを話者推定タスク用に学習する手法 [9, 13] や，GPT-3.5 や GPT-4 [14] で入力するプロンプトを調節することで話者名を識別する手法 [8, 15] などがある．

一方で，特にコンテキストウィンドウのサイズに関する制約が依然として存在する．Albert ら [16] は，LLaMA-3 [17] がコンテキストウィンドウを拡大し，PDNC における精度を改善したことを示したが，その評価はモデルや言語の範囲が制約されており，不完全なものであった．

3 提案手法

話者推定　話者推定は，物語の文章に登場する登場人物やキャラクターの中で，各台詞が誰によって発話されたものなのかを推定するタスクである．我々は，物語に登場する人物で文章上で主に記述される名前（例：玄徳）とその言い換え表現（例：玄徳，劉備，蜀王）の 2 種類に対してアノテーションを実施した．

表 1: データセットごとのトークン数と話者数。本データセットは，LLaMA-2 トークナイザを基準とした台詞前後のコンテキスト長（1,024 トークン）に基づいて抽出されている．`book_id=052409` は序章を表し，「三国志演義」の物語の序盤を描いている．桃園の巻（`book_id=052410`）から三国志の最終章（`book_id=052420`）までは，データセットは物語の時系列の進行に従っている．`book_id=052410` は開発（dev）データとして利用され，すべて人手でアノテーションされた．一方，`book_id=052411--052420` は評価（eval）データとして使用され，LLM によって最初に生成されたラベルを人手で修正した．

book_id	タイトル	トークン数	台詞数	skip	話者数
除外データ					
052409	序	1,866	0	2	0
開発用データ（dev）：全て人手のアノテーション					
052410	桃園の巻	195,226	1,686	70	113
評価用データ（eval）：LLM ラベル + 人手修正					
052411	群星の巻	195,589	1,662	108	157
052412	草莽の巻	193,973	1,649	129	136
052413	臣道の巻	201,042	1,616	82	123
052414	孔明の巻	205,799	1,461	89	159
052415	赤壁の巻	209,759	1,532	88	117
052416	望蜀の巻	204,514	1,598	83	153
052417	図南の巻	222,992	1,433	95	171
052418	出師の巻	249,258	1,426	96	186
052419	五丈原の巻	223,710	1,308	130	122
052420	篇外余録	27,050	40	40	26
合計		2,130,778	15,411	1,012	1,463

プロンプト調節　効率良く高品質なデータセットを作成するために，我々は少数の開発用データセットに対して人手でアノテーションし，話者推定を行う指示に従うように LLM への入力用プロンプトの調節を実施した．また調整されたプロンプトでは，チャットテンプレート[1)]と回答方法を指示するための少数事例を入力に含めた．

4 データセット構築

以下のステップでデータセット構築を行った．

STEP 1: 前処理　青空文庫[2)]の「三国志」から台詞を正規表現のルールベースで収集し，LLaMA-2 トークナイザでトークン化した上で，各データに対して台詞の前後 1,024 トークンを周辺コンテキストとして抽出した．これにより，合計 16,423 件の台詞とそのコンテキストのペアからなるデータセットを

1) https://github.com/chujiezheng/chat_templates
2) https://www.aozora.gr.jp/

構築した．本データセットは全 11 冊から構成され，book_id=052410 を開発用データ，book_id=052411 から 052420 を評価用データとした．

STEP 2: LLM による話者推定　抽出した対話文中の各台詞に対して，LLM を用いて話者ラベルを付与した．データセット構築では，開発時に性能が高かった `LLaMA-3-70B-Instruct` の出力値を採用した．

STEP 3: 人手修正　アノテーション規則に基づき（Appendix A 参照），話者名を手動で修正し，識別されたラベルの約 10%を調整した．ただし，コンテキスト内で話者名に対応する語彙が欠落しているケースや，1 つの対話に複数の話者が混在する場合は除外した．この作業で 1,011 件を削除し，最終的に 15,412 件のデータセットを構築した[3)]．GPU の推論には NVIDIA RTX A6000 を 30 時間使用した．

この手法により，データセット構築に要する時間を大幅に短縮した．当初 1,500 件のデータを手動でアノテーションするには 10 時間を要していたが，修正作業に特化することで 1,500 件あたり 3.5 時間に短縮された．表 1 は，各 book_id ごとのトークン数（LLaMA-2 トークナイザー基準），台詞数，スキップした件数，および物語に登場する話者の数を示している．アノテーションされた話者数の内，重複を除いた数は 856 名であった．

人手評価　アノテーションの品質を検証するため，評価用データセットからランダムに抽出した 100 件のサンプルを利用して，データセット作成者とは独立した 3 名のアノテータに評価を依頼した．人手評価では，作成したデータセットの話者名に対して「適切」，「不適切」，または「判断不能」の 3 択から選び，「適切」に関する一致率を算出した．その結果，高い整合性を示し，2 名は 0.97，1 名は 0.96 の一致率を達成した（付録 B 参照）．

5　実験

我々は構築したデータセット上で LLM が持つ話者推定能力を評価し，データセット構築を効率良く実施するための LLM 分析を実施した．

5.1　プロンプト

表 2 に示すように，プロンプトにはチャットベースのテンプレートを用いて，LLM に入力する形で話者推定タスクを実施した．

3)　https://huggingface.co/datasets/satoshi-2000/romance_of_the_three_kingdoms

表 2: 話者推定用のプロンプト．`{Example Context}` と `{Context}` は物語として入力するコンテキスト，`{Line}` は話者推定の対象となるセリフを示している．

Role	Content
user	次の物語（# Example Story）中のセリフに対して、誰が発言したのかを推測して、話者の名前のみを答えてください。
assistant	了解しました。以下の物語とセリフに基づいて答えを提供します。
user	# Example Story `{Example Context}`
assistant	物語を確認しました。では、セリフごとに誰が発言したのかを答えます。
user	次の発話は誰が発言しましたか？
assistant	セリフを教えてください。
user	おーい
assistant	漁夫
user	――そこの若い者ウ。なにを見ているんだい。いくら待っていても、そこは渡し舟の着く所じゃないぞ
assistant	漁夫
user	ありがとう
assistant	青年
user	おい、おい、旅の者
assistant	百姓
user	――そんな所で、今朝からなにを待っているんだね。このごろは、黄布賊とかいう悪徒が立ち廻るからな。役人衆に怪しまれるぞよ
assistant	百姓
user	同様にして、次の物語（# Target Story）中のセリフに対して、誰が発言したのかを推測して、話者の名前のみを答えてください。
assistant	了解しました。以下の物語とセリフに基づいて答えを提供します。
user	# Target Story `{Context}`
assistant	物語を確認しました。では、セリフごとに誰が発言したのかを答えます。
user	次の発話は誰が発言しましたか？
assistant	セリフを教えてください。
user	`{Line}`

5.2　モデル

効率的なデータセット構築に向けた比較のため，汎用的に性能の高い LLaMA-3 [17] を基準とし，さらに日本語データで追加学習した Swallow-3[4)]，ELYZA-JP-8B[5)]，および LLaMA-3-youko-8B[6)] を選択した．より広範なモデル評価のため，Mistral 7B[7)] および日本語データで学習した RakutenAI-7B[8)] も

4)　https://huggingface.co/collections/tokyotech-llm/llama-3-swallow-667e904b34c4cc3d4e48e085
5)　https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
6)　rinna/llama-3-youko-8b
7)　https://huggingface.co/mistralai/Mistral-7B-v0.1
8)　https://huggingface.co/Rakuten/RakutenAI-7B

表 3: 全体的な話者推定の識別精度：主要な話者名として付与されたラベルに対する各モデルの完全一致率，部分一致率，編集距離，BERTScore (F1) の開発用データ（dev）および評価用データ（eval）における各指標の平均値を示す．

モデル	完全一致率↑		部分一致率↑		編集距離↓		BERTScore (F1)↑	
	dev	eval	dev	eval	dev	eval	dev	eval
Swallow-3-8B	0.219	0.226	0.520	0.547	7.751	7.840	0.792	0.804
Swallow-3-8B-Instruct	0.465	0.573	0.794	0.786	1.543	1.243	0.888	0.915
Swallow-3-70B	**0.803**	**0.847**	**0.864**	**0.885**	**0.446**	**0.354**	**0.959**	**0.968**
Swallow-3-70B-Instruct	**0.802**	**0.845**	**0.895**	**0.901**	**0.476**	**0.381**	**0.958**	**0.968**
Karakuri-8x7B	0.658	0.698	0.735	0.730	0.845	0.731	0.923	0.934
Mistral-7B	0.000	0.000	0.469	0.476	10.423	11.281	0.676	0.678
RakutenAI-7B	0.138	0.118	0.725	0.700	6.837	6.828	0.772	0.770
ELYZA-JP-8B	0.483	0.533	0.530	0.568	1.432	1.367	0.706	0.740
llama-3-youko-8B	0.345	0.299	0.563	0.527	5.852	6.734	0.812	0.795
LLaMA-3-8B-Instruct	0.537	0.561	0.648	0.649	2.705	2.553	0.877	0.889
LLaMA-3-70B-Instruct	**0.781**	**0.830**	**0.863**	**0.888**	**0.620**	**0.425**	**0.950**	**0.965**
CALM-3-22B	0.580	0.516	0.664	0.596	4.240	4.761	0.879	0.864

評価に含めた．また，学習データの構成が精度へ与える影響を評価するため，主に日本語データで学習された CALM-3-22B[9]や，Mixture of Experts [18] のモデルをベースに日本語データで学習した Karakuri-8x7B[10]も分析対象とした．

5.3 評価指標

完全一致率 生成テキスト内で識別された話者とアノテーションにおける話者が完全に一致する割合を測定するものであり，従来研究 [16] で一般的に用いられている．

部分一致率 LLM が生成したテキストにはノイズを含む場合があるため，話者名の主要部分が部分的に一致するかを判定する．

編集距離 [19] 編集距離は，文字挿入・削除・置換の回数により 2 つの文字列間の類似度を測定する指標である．

BERTScore [20] BERTScore は埋め込み表現に基づいて類似性を評価する指標であり，表層上の表現が異なっていても意味が変わらない場合を評価することができる．

5.4 実験結果と考察

総合性能 表 3 に，各モデルの話者推定精度を示す．開発（book_id=052410）および評価（book_id=052411–052420）の部分一致率で約 90% の精度が得られ，LLM は話者推定において堅牢な性能を示した．最高精度を達成した `Swallow-3-70B-Instruct` は，LLaMA-3 をベースとし，日本語データでの継続的事前学習を経てインストラクションチューニングを施したモデルである．また，オリジナルの LLaMA-3 モデルはこれに次ぐ 2 番目の成績を示した．さらに，`Swallow-3-8B-Instruct` は `Swallow-3-8B` より 5%精度が向上しており，インストラクションチューニングの有効性が示された．

これらの結果は，大規模モデルと高品質なデータセットによる追加学習の組み合わせが話者推定を正確に実施する場合でも重要であることを示している．タスク対象の言語のデータを用いた継続的事前学習とインストラクションチューニングをすることで，データセットを高精度で構築するための人手修正コストを削減することが期待される．

6 おわりに

本研究では，青空文庫版「三国志」から 15,412 件の発話に対して，LLM を用いて話者ラベル付けを行い，人間による修正を加えることで効率的に物語分析用データセットを構築する手法を示した．LLaMA-3 を用いることで約 90 %と高精度で正確な話者名を推測することが可能となった．このアプローチにより，人手作業の大幅な削減と高品質なアノテーションの両立が可能であることを示した．

今後は，話者の呼称先や引用タイプも含めたデータセットの拡張と，大規模で複雑な物語への LLM の対応力を検証予定である．これにより，話者推定に留まらず，物語分析におけるキャラクター間関係，感情傾向，多言語・異文化比較，教育・言語学習への応用など，多面的な展開が期待される．

9） https://huggingface.co/cyberagent/calm3-22b-chat
10） https://huggingface.co/karakuri-ai/karakuri-lm-8x7b-chat-v0.1

参考文献

[1] Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 298–311, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

[3] David Elson and Kathleen McKeown. Automatic attribution of quoted speech in literary narrative. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 24, No. 1, pp. 1013–1019, Jul. 2010.

[4] Hua He, Denilson Barbosa, and Grzegorz Kondrak. Identification of speakers in novels. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1312–1320, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. A two-stage sieve approach for quote attribution. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 460–470, Valencia, Spain, April 2017. Association for Computational Linguistics.

[6] Jia-Xiang Chen, Zhen-Hua Ling, and Li-Rong Dai. A Chinese Dataset for Identifying Speakers in Novels. In **Proc. Interspeech 2019**, pp. 1561–1565, 2019.

[7] Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5838–5848, Marseille, France, June 2022. European Language Resources Association.

[8] Yuki Zenimoto, Shinzan Komata, and Takehito Utsuro. Large scale evaluation of end-to-end pipeline of speaker to dialogue attribution in Japanese novels. In Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Si Chen, Yu-Yin Hsu, Emmanuele Chersoni, Pranav A, Winnie Huiheng Zeng, Bo Peng, Yuxi Li, and Junlin Li, editors, **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 12–23, Hong Kong, China, December 2023. Association for Computational Linguistics.

[9] 石川和樹, 小川浩平, 佐藤理史. 口調エンコーダを用いた小説発話の話者推定. 自然言語処理, Vol. 31, No. 3, pp. 894–934, 2024.

[10] 杜宇龍. 小説からの対話コーパスの自動構築. 2019.

[11] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey, 2024.

[12] David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian mixed effects model of literary character. In Kristina Toutanova and Hua Wu, editors, **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 370–379, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[13] Carolina Cuesta-Lazaro, Animesh Prasad, and Trevor Wood. What does the sea say to the shore? a BERT based DST style approach for speaker to dialogue attribution in novels. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5820–5829, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[15] Zhenlin Su, Liyan Xu, Jin Xu, Jiangnan Li, and Mingdu Huangfu. Sig: Speaker identification in literature via prompt-based generation. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, No. 17, pp. 19035–19043, Mar. 2024.

[16] Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. A realistic evaluation of llms for quotation attribution in literary texts: A case study of llama3, 2024.

[17] Meta. The llama 3 herd of models, 2024.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[19] Vladimir I Levenshtein, et al. Binary codes capable of correcting deletions, insertions, and reversals. In **Soviet physics doklady**, Vol. 10, pp. 707–710. Soviet Union, 1966.

[20] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

A　アノテーションルール

1. 物語中で用いられた登場人物名のうち，最も短い構成要素を正解ラベルとする（例：「劉備玄徳」は「玄徳」を使用）.
2. 複数候補がある場合，文脈中に存在すれば与名を優先する.
3. 対話でない場合は「Unknown」とする（例：人物一覧，語り手，書名）.
4. 1 つの発話に複数の話者が含まれる場合は「Unknown」とする（例：関羽と張飛が同時に発話している）.
5. 各発話は前後 1,024 トークンを含むコンテキストと共に処理し，その範囲で確認できる名前のみをアノテーション対象とする（トークン数は LLaMA-2 トークナイザで計測）.
6. 単一人物に関して複数の名称が文脈中にある場合，最も適切なものを主要な呼称とし，他を言い換え表現とする.

B　データセットの人手評価

本データセットのアノテーションでは，物語を熟読する必要があるため大量の時間を要し，データセット全体に対して人手評価を実施するのは困難である．そこで本研究では，品質確認のために評価データから無作為に 100 件抽出し，独立した 3 名のアノテータにより「適切」「不適切」「判断不能」の 3 区分で一致率を評価した．一致率を計算した結果，アノテーションデータに対して妥当なラベルであると判断されたデータの割合は，2 名が 0.97，1 名が 0.96 と高かったことから，本データセットの品質の高さが示唆された．また，人手評価のためのアノテーション作業はアノテータ 1 人あたり平均 2 時間を要し，時給 1,000 円で日本人大学院生 3 名に対して実施した.

C　推測例

C.1　長い対話における誤り

モデルは最後の「なにが？」を妻の発話と誤認したが，実際には楊彪が発したものである．話者交替が頻繁な長い対話では，このような誤りが生じやすい傾向が観測された.

コンテキスト（抜粋）：

楊彪は秘策を胸にねりながら、わが邸へ帰って行った。帰るとすぐ、彼は妻の室へはいって、「どうだな。この頃は、郭しの令夫人とも、時々お目にかかるかね。……おまえたち奥さん連ばかりで、よく色々な会があるとのことだが」二 楊彪の妻は怪しんで、良人を揶揄した。「あなた。どうしたんですか、いったい今日は」「なにが？」

C.2　ナレーションの識別

ナレーションによる補足説明「江東の地」に対して，LLM（LLaMA-70B-Instruct）は登場人物による発話ではなく，「ナレーション」の台詞であると正しく判定していた.

コンテキスト（抜粋）：

呉は，大江の流れに沿うて，「江東の地」と称われている。

C.3　沈黙している発話者の推定

LLM（LLaMA-70B-Instruct）は「……」を貂蝉による沈黙と適切に話者を推定していた.

コンテキスト（抜粋）：

貂蝉は、さわぐ色もなく、すぐ答えた。「はい。大人のおたのみなら、いつでもこの生命は捧げます」王允は、座を正して、「では、おまえの真心を見込んで頼みたいことがあるが」「なんですか」「董卓を殺さねばならん」「……」

C.4　複数話者が存在する場合の対処

LLM（LLaMA-70B-Instruct）は，台詞「人生の快、ここに尽くる」が関羽および張飛の 2 名によって発話されていることを正確に識別していた.

コンテキスト（抜粋）：

「人生の快、ここに尽くる」関羽、張飛がいうと、「何でこれに尽きよう。これからである」と、玄徳はいった。

C.5　事前知識への依存

LLM（ELYZA-JP-8B）は，入力されたコンテキストを削除しても，関羽の発話であると正しく推定した．このことから，「玄徳」という人名に基づいて連想し，事前に学習された知識に影響を受けた可能性が高いと考えられる.

コンテキスト（全体）：

「玄徳様、ふたりの熱望です。ご承知くださるまいか」