LLM の生成テキストの真偽検証のための日本語言説分解データセットの構築と評価

政野 美和 [1,2]　欅 リベカ [3]　欅 惇志 [1]　清丸 寛一 [2]　中山 功太 [2]
堀尾 海斗 [4]　源 怜維 [2,4]　橘 秀幸 [2]　河原 大輔 [2,4]
[1] 一橋大学 ソーシャル・データサイエンス学部
[2] 国立情報学研究所 大規模言語モデル研究開発センター
[3] 東京工科大学コンピュータサイエンス学部　[4] 早稲田大学理工学術院
{a.keyaki@r,5123053k@g}.hit-u.ac.jp　{kiyomaru,nakayama,h_tachibana}@nii.ac.jp
keyakirbk@stf.teu.ac.jp　{kakakakakakaito,ray}@akane.waseda.jp　dkw@waseda.jp

概要

LLM の生成テキストは誤情報を含むことがあるため，情報の真偽を検証するシステムの開発は喫緊の課題である．言説単位の真偽検証システムでは，生成テキストを独立した最小粒度の情報である言説 (claim) に分解して，言説ごとに情報の真偽を判定する．言説単位での真偽検証を行うことで，真偽検証結果に対する説明性が向上する．高性能な言説分解手法の開発・評価のため，本研究では日本語言説分解データセットの構築に取り組んだ．

1　はじめに

大規模言語モデル（Large Language Model; LLM）が社会において広く利用され始めているが，LLM の生成テキストにはハルシネーション（hallucination）と呼ばれる，もっともらしく見える誤情報が含まれることがある [1]．こうした背景から，LLM の生成するテキストに対して，情報の真偽を検証するシステム（真偽検証システム）の研究・開発が進んでいる [2, 3, 4, 5, 6]．我々の研究グループでも LLM の生成テキストに対する日本語真偽検証システムの開発に取り組んでいる [7, 8]．これらにより，LLM の信頼性が高まることが期待される．

現在主流の真偽検証システムの概要を図 1 に示す．まず，(1) 言説分解 (claim decomposition) では，入力テキスト（LLM の生成テキスト）を，一つの物事に関する性質や関係を表す独立した最小粒度の情報単位である言説 (claim) に分解する．次に，(2) 根拠検索 (evidence retrieval) では，言説をクエリとして検索を行い，関連する根拠テキストを取得する．最後に，(3) 真偽判定 (verdict prediction) では，取得された言説-根拠テキストのペアに対して根拠テキストが言説の内容を支持するかを判定する．

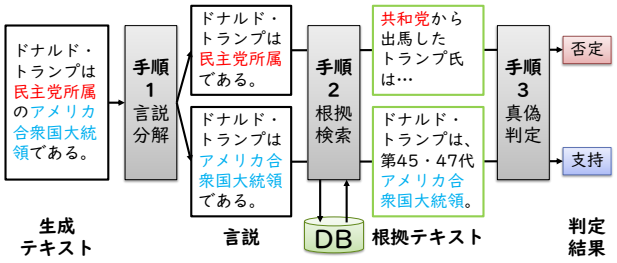

図 1　真偽検証システムの概要図

言説単位で真偽検証を行うことで，細かい粒度での誤情報を検出することができるため，真偽検証結果に対する説明性が向上する．言説分解の品質は真偽検証の性能に大きな影響を与えることが報告されている [9] ため，高品質な言説分解は高性能な真偽検証システムの開発において非常に重要である．

本研究では，日本語の真偽検証システムにおける言説分解性能の改善のため，生成テキストとその言説セットが含まれる言説分解データセットを構築する．言説分解データセットの開発により，(i) 言説分解手法の定量評価，(ii) 言説分解に存在する課題を明らかにするための分析，(iii) 評価と分析に基づく言説分解手法の改善が可能となる．

本論文では，著者らが過去に作成した言説分解ガイドライン [7] を用いて構築したデータセットについて報告する．定量的な評価の結果，構築したデータセットは高品質であることが確認された．さらに，本データセットを用いた，プロンプトによる言説分解の実験の結果，プロンプトにガイドラインを加えることで性能が改善することが示された．

2 関連研究

文献 [9] は，言説を粒度とした事実性に関する評価タスク（生成テキストの事実精度評価 [10]，含意関係認識 [11]，真偽検証 [5, 6]）において，言説分解の品質は各タスクの性能に大きな影響を与えることを実験的に確認した．これは，言説分解によって言説の個数や個々の言説の持つ情報が決まるため，言説分解におけるエラーは事実性評価にも影響を及ぼすためである．なお，文献 [9] では，高品質な言説分解のためには下記の三つの性質を満たす必要があると報告している．

網羅性 (coverage) 分解で作成された言説が全体として元の生成テキストのすべての部分を含む

整合性 (coherence) 分解で作成された個々の言説が元の生成テキストに忠実である（含意される）

原子性 (atomicity) 言説が可能な限り原子的である（粒度が細かい）

3 データセット構築

言説分解データセットは，LLM の生成テキスト，生成テキストから言説分解により作成された言説，個々の言説に対する言説分類のラベルから構成される．以下は生成テキスト (1) に対する 2 つの言説 (2) と，個々の言説の分類ラベルの例である．

(1) ドナルド・トランプは民主党所属のアメリカ合衆国大統領である。

(2) 1. ドナルド・トランプは民主党所属である。
言説分類ラベル：Check-worthy

2. ドナルド・トランプはアメリカ合衆国大統領である。
言説分類ラベル：Check-worthy

3.1 言説分解ガイドライン

言説分解では，生成テキストを可能な限り細かい粒度に分解することを基本としつつ，言説が生成テキストから含意される範囲で分解することに留意しなければならない．そのため，データセット構築に先立って言説分解ガイドライン [7] を作成した．

ガイドラインではまず以下の三原則を定めた．

原則 1 各言説は前後の文脈なしに解釈できるようにする

原則 2 分解された言説から元の情報を復元できるようにする

原則 3 なるべく元のテキストと同じ表現を用いるようにする

さらに，個別の詳細な分解ルールにおいて，並列構造や関係節等の分解可能な構造に対する分解手順を定め，意味的に分解不可能な例外的事例を示した．加えて，生成テキストの主題と見なせ，検索における検出力が高いことが期待される表現（焦点語と呼ぶ）が主語となるように分解するなど，後段の処理を考慮した表現への修正手順を示した．

分解して得られた言説のうち，客観的な事実や評価に関する言説には真偽検証に値することを表す Check-worthy のラベルを付与し，そうでない言説には Not check-worthy を付与することとした．

3.2 LLM によるテキスト生成

後述の二つのデータセットに対して，LLM (`llm-jp/llm-jp-3-13b-instruct`) を用いてテキストを生成した．

AIO: AI 王 Version 2.0 開発用データ[1] はクイズ問題を題材とした日本語質問応答データセットである．AIO は質問応答データセットである性質上，事実性を問う質問が多数含まれている．

CBA: LLM-jp Chatbot Arena Conversations Dataset[2] はさまざまな LLM を公平に比較・評価するためのオープンなオンラインプラットフォームである．CBA には事実性を問う質問以外にもユーザの多様な入力が含まれ，より現実的なシナリオを想定した対話データセットである．

テキスト生成の詳細な手順は付録 A に掲載する．

3.3 人手による言説分解とラベル付与

言説分解と言説分類ラベル付与は，アノテーションの経験を持つ非専門家に依頼した．作業者には著者らが作成したサンプルの提供を行い，また，作業前にガイドラインを熟読してもらった．作業体制は，1 名の作業者が言説分解と言説分類ラベル付けを行い，1 名のチェッカーがその作業結果を確認して，不十分な点があれば 2 名で協議を行った．その際，判断に迷ったケースに対しては注釈を付与した．作業中の不明点には著者らが随時回答し，必要

1) https://sites.google.com/view/project-aio/dataset
2) https://huggingface.co/datasets/llm-jp/llm-jp-chatbot-arena-conversations

表 1　データセットの統計量

	AIO	CBA
生成テキスト件数	420	97
平均文数	2.00	15.32
平均言説数	5.55	32.11
言説数の標準偏差	2.89	20.02
言説の平均語数	15.71	17.42
言説の平均内容語数	6.90	8.23
check-worthy な言説の比率	0.98	0.87

に応じてガイドラインの更新を行った．

著者らによる検収作業において，CBA の 3 件の生成テキストは適切に言説分解することが困難であると判明したため，データセットから除外した．そのため，最終的に AIO は 420 件，CBA は 97 件の生成テキストに対しての言説と言説分類ラベルが付与された．データセットの概要を表 1 に示す．

構築したデータセットを開発用データとテスト用データを 1 対 1 の比率で分割した．AIO の開発データ・テストデータはそれぞれ 210 件で，CBA ではそれぞれ 48 件である．なお，AIO の開発データには，ガイドライン中でルールの説明に用いた生成テキストが含まれる．それ以外の生成テキストは比率に応じてランダムに割り振られた．

3.4　データセットの品質評価

構築したデータセットの品質評価を行う．言説分解の品質評価のための指標 Decompscore [9] を拡張して，下記の NormalizedDecompScore (NDS) によって言説分解の品質を計算する．

$$NDS = \frac{1}{N}\sum_{t=1}^{N}\frac{C_t^{\text{entailed}}}{C_t} \tag{1}$$

N は評価対象の生成テキストの数，C_t は生成テキスト t に含まれる言説の数，C_t^{entailed} は生成テキスト t から含意される言説の数とする．そのため，NDS は生成テキストに含意される言説の比率を表し，言説分解の品質を表す．言説が元のテキストから含意されるかの判定は `Qwen/Qwen3-32B`[3] を用いた．

AIO: AIO の NDS スコアは 0.980 と極めて高い値を達成した．このことから，AIO の品質の高さが確認された．作成された言説分解結果と著者らが独立に作成した言説分解結果を比較したところ，複雑な構造を持つ生成テキストに対しても概ね分解結果が一致した．その一方で，並列構造の解釈，焦点語の選択，作業者の知識などの差異によって作業者間で揺れが生じた分解も存在した．揺れの軽減が期待されるルールの導入に関して引き続き検討を行う．

CBA: CBA もまた NDS スコアは 0.975 と極めて高い値を達成した．言説分解において文脈依存性を解消する際に，ユーザの入力情報を使わないと適切に文脈を補完できないケースが存在した．今後の課題として，必要に応じて入力からの情報の補完を行うよう，作業手順を修正する必要がある．

4　プロンプトによる言説分解

本節では，構築したデータセットを用いてプロンプトベースの言説分解の性能評価を行う．紙面の制約上，主に AIO に対する実験結果を報告する．

4.1　言説分解手法

評価対象の手法を述べる．まず，タスクの指示と言説の定義のみを含めたプロンプトをベースプロンプト (base) とする．提案手法のプロンプトの一つとして，言説分解ガイドラインを含むプロンプト (guideline) を作成する．`guideline` は，`base` に加えて，言説分解ルール，言説分解の例，例に対する解説から構成される．なお，`guideline` には合計 12 件の言説分解の例が含まれる[4]．

また，プロンプト構築の主要な手法である，`few-shot` の性能も検証する．shot 数の変動による性能への影響を評価するため，shot 数は，2, 4, 6, 8 を選択した．事例は開発データから抽出した．なお，事例の選択は，多様な性質の事例が含まれるように注意深く行った．また，比較のため，`guideline` に含まれる 12 件の言説分解の例を few-shot として与える設定 (`guideline-based shot`) も評価した．

さらに，`guideline` と `8-shot` の組合せ (`guideline+8-shot`) も検証する．

最後に，既存の言説分解手法との比較を行うため，`FactScore` [10] と `R-ND` [9] のプロンプトに対する評価も行った．これらのプロンプトでは先行研究 [9] にならって 8-shot を採用している．用いる 8-shot は提案手法と同じ事例である．

4.2　評価指標

プロンプトによって分解された言説（予測言説）と，構築されたデータセットに含まれる言説（正

3)　https://huggingface.co/Qwen/Qwen3-32B

4)　指示の単純化の観点から一部の例については生成テキストを部分的に使用しているため，実際の生成テキストに対する言説分解時よりも分解後の言説数は少なくなる傾向がある．

表 2　プロンプトベースの言説分解の性能 (AIO)

プロンプト	shot 数	Exact match			Fuzzy match			分解率
		Prec.	Rec.	F1	Prec.	Rec.	F1	
base	0	0.009	0.007	0.008	0.412	0.358	0.374	0.797
+2-shot	2	0.150	0.119	0.130	0.529	0.416	0.457	0.720
+4-shot	4	0.197	0.169	0.179	0.560	0.479	0.507	0.796
+6-shot	6	0.184	0.163	0.171	0.562	0.500	0.520	**0.834**
+8-shot	8	0.202	0.173	0.184	0.561	0.486	0.512	0.804
+guideline-based shot	12	0.156	0.132	0.141	0.545	0.442	0.480	0.724
guideline	1	0.159	0.129	0.140	0.509	0.399	0.439	0.681
+8-shot	8	**0.223**	**0.192**	**0.204**	**0.587**	**0.504**	**0.534**	0.807
FactScore [10]	8	0.193	0.169	0.177	0.540	0.472	0.494	0.806
R-ND [9]	8	0.195	0.169	0.179	0.545	0.476	0.499	0.809

解言説）を比較することで性能評価を行う．言説のマッチング手法および評価尺度は，言説単位の NLI [11] にならった．

Exact match　完全一致する予測言説と正解言説のペアを「正しいマッチング」として扱う．

Fuzzy match　予測言説セット $\{p_i\}$ と正解言説セット $\{g_j\}$ の類似度を評価するため，まず語レベルの Jaccard 係数 $J_{ij} = |p_i \cap g_j|/|p_i \cup g_j|$ の類似度行列を算出する．次に，ハンガリアン法 [12] を用いて，予測言説と正解言説の間の最適な二部グラフのマッチングを判定する．予測言説と正解言説の Jaccard 係数が閾値 θ を上回る場合には「正しいマッチング」として扱う．文献 [11] にならって，閾値 θ に 0.8 を設定する．なお，内容語（名詞，動詞，副詞，形容詞）のみを対象とした場合の性能を報告する．

上記のマッチング手法それぞれに対して，適合率，再現率，F1 のマクロ平均を算出する．適合率は予測言説のうち「正しいマッチング」と判定された言説の割合，再現率は正解言説のうち「正しいマッチング」と判定された言説の割合で算出される．また，各手法の性能は 5 回実行した際の平均値を報告する．

4.3　言説分解の実験結果

実験結果を表 2 に掲載する．few-shot は概ね例の数が多くなるほど性能が向上した．ガイドライン中の例を用いた gudeline-based shot は，より少数の例を用いた few-shot よりも性能が低かった．この結果は，高性能な言説分解には高品質な例を用いることが重要であるという知見 [9] とも整合する．guideline は性能を改善させたものの，その効果が限定的であった．対して，guideline+8-shot は全評価手法の中で最も高性能を達成した．さらに，FactScore と R-ND は 8-shot と同等の性能を示した．

いずれの手法においても，適合率よりも再現率が低いという結果が得られた．そのため，正解言説の件数に対する予測言説の件数（分解率）を算出した．その結果，全体的に 7, 8 割程度の分解率であることが判明した．分解率と性能には概ね相関があったが，最も分解率の高い 6-shot よりも guideline+8-shot の方が性能が高かった．このことは，言説を可能な限り細かく分解することと，適切な言説への分解の両立の必要性を示している．

例えば，「この海峡は黒海とマルマラ海を結び」という生成テキストにおいて，作業者は「黒海とマルマラ海」を意味的に分解不可能な並列構造と判断したのに対して，guideline+8-shot では適切に「黒海とマルマラ海」を分解しなかった．

CBA においても，guideline+8-shot が最も高い性能を示した．また，性能や分解率は全般的に AIO よりも低く（Fuzzy match F1 は 0.431），shot 数を増やしても性能改善が限定的という結果が得られた．このことから，CBA はより難易度の高いデータセットとなっているといえる．詳細な実験結果は付録 B に記載する．

また，付録 C のマッチング結果の分析の結果，閾値 θ に 0.8 を設定することで正しいマッチングとそれ以外を適切に判別できていると考えられる．

5　おわりに

本研究では，LLM の生成テキストの真偽検証システム構築のための日本語言説分解データセットを構築した．構築されたデータセットは高品質であることが確認された．また，高品質な事例やガイドラインを追加することでプロンプトベースの言説分解の性能向上が確認された．

謝辞

本研究成果の一部は，データ活用社会創成プラットフォーム mdx を利用して得られたものである．また，本研究の一部は，JSPS 科研費（基盤研究 (B)（課題番号: 23H03686, 25K03178），基盤研究 (C)（課題番号: 24K15066)），令和 7 年度次世代人工知能技術等研究開発拠点形成事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」，株式会社デンソーアイティーラボラトリとの共同研究の支援による．ここに記して謝意を表す．

参考文献

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Trans. Inf. Syst.**, Vol. 43, No. 2, January 2025.

[2] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 5430–5443, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. Automated fact-checking: A survey. **Language and Linguistics Compass**, Vol. 15, No. 10, p. e12438, 2021.

[4] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 178–206, 2022.

[5] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 7561–7583, Singapore, December 2023. Association for Computational Linguistics.

[6] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 14199–14230, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] 政野美和, 欅リベカ, 欅惇志, 清丸寛一, 中山功太, 堀尾海斗, 源怜維, 橘秀幸, 河原大輔. LLM の生成テキストの真偽検証のための日本語言説分解データセットの構築. 第 265 回 自然言語処理研究発表会, 2025.

[8] 政野美和, 清丸寛一, 欅惇志, 堀尾海斗, 源怜維, 欅リベカ, 中山功太, 橘秀幸, 河原大輔. LLM の生成テキストの真偽検証のための日本語真偽判定データセットの構築. 言語処理学会第 32 回年次大会 (NLP2026), 2026.

[9] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In Danushka Bollegala and Vered Shwartz, editors, **Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)**, pp. 153–175, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[10] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

[11] Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. PropSegmEnt: A large-scale corpus for proposition-level segmentation and entailment recognition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8874–8893, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Harold William Kuhn. The Hungarian method for the assignment problem. **Naval Research Logistics Quarterly**, Vol. 2, No. 1-2, pp. 83–97, 1955.

表 3　プロンプトベースの言説分解の性能 (CBA)

プロンプト	shot 数	Exact match			Fuzzy match			分解率
		Prec.	Rec.	F1	Prec.	Rec.	F1	
base	0	0.007	0.004	0.005	0.452	0.286	0.337	0.492
+2-shot	2	0.102	0.059	0.073	0.505	0.304	0.370	0.492
+4-shot	4	0.127	0.083	0.098	0.523	0.337	0.400	0.531
+6-shot	6	0.110	0.069	0.083	0.484	0.311	0.369	0.538
+8-shot	8	0.129	0.082	0.098	0.528	0.337	0.401	0.536
+guideline-based shot	12	0.129	0.082	0.097	0.492	0.299	0.356	0.475
guideline	1	**0.146**	**0.099**	**0.115**	0.540	0.366	0.423	0.562
+8-shot	8	0.145	0.098	0.114	**0.552**	**0.366**	**0.431**	**0.565**
FactScore [10]	8	0.117	0.074	0.089	0.499	0.325	0.381	0.529
R-ND [9]	8	0.123	0.078	0.093	0.508	0.332	0.389	0.535

A　テキスト生成

AIO テキスト生成: データセットの開発用データ 1,000 件の質問に対する解答を LLM に生成させた．そのうち，有効な解答は 864 件であった．著しく短い解答は生成テキストから文脈が欠落しているケースが頻繁に起こっていた（解答例：「「e スポーツ」と言います。」）ため，24 文字以下の解答は除外した．また，第一弾のバージョンでは 151 文字以上の解答も除外した．その結果，解答は 420 件収集された．

CBA テキスト生成: データセット中の 920 件の入力に対する解答を LLM に生成させた．そのうち，有効な解答は 632 件であった．CBA の中には挨拶や数値計算のような，真偽検証の対象外となる入力も多数含まれていた（回答例：「あけましておめでとうございます！」「USD/JPY が 123 円のとき$100 は何円?」）．そのため，複数の LLM（Qwen3-32B と gemma-3-27b-it）によって check-worthy な言説を含むと判定された生成テキスト 267 件を抽出した．そのうちの 100 件の生成テキストをランダムにサンプリングして用いることとした．

B　CBA の言説分解性能評価

表 3 に CBA に対する実験結果を示す．なお，few-shot の例は AIO における実験と共通である．AIO の実験においては，guideline よりも高性能な few-shot が存在したが，CBA では guideline は一貫して few-shot の性能を上回った．このことから，AIO と CBA では性質が異なり，few-shot 例を共有することは最適ではないことを示唆している．そのような状況においても，ガイドラインを与えることで言説分解性能を改善できることが明らかになった．

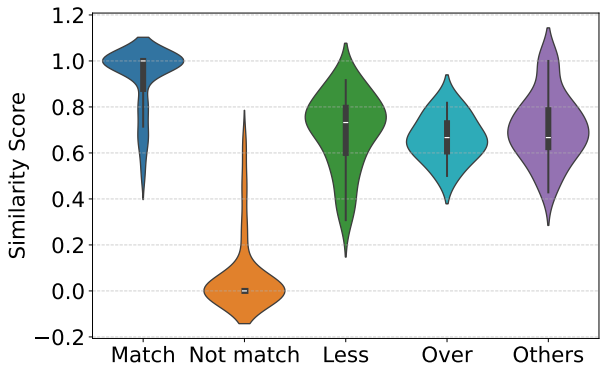

図 2　類似度スコア分布

C　言説分解マッチング結果の分析

4.3 節の言説分解のマッチング結果の分析のために，guideline+8-shot のマッチング結果に対して，著者らによるアノテーションを行った．アノテーションのラベルは，下記の通りである．

Match　マッチング結果が正しい (53%).
Not match　マッチング結果が正しくない (20%).
Less　予測言説の分解が正解言説より粗い (11%).
Over　予測言説の分解が正解言説より細かい (3%).
Others　焦点語の不一致や，言説分解ルールへの違反

アノテーション対象は，生成テキスト 50 件分（言説 295 件）である．

アノテーション結果を図 2 に示す．*Match* の類似度スコアの中央値は 1.0 付近であり，誤ったマッチング *Not match*, *Over*, *Less*, *Others* の類似度スコアの中央値は 0.8 未満である．そのため，閾値 θ に 0.8 を設定することで正しいマッチングとそれ以外を適切に判別できていると考えられる．