大規模言語モデルを用いたニュース類似度の算出

井本稔也
Japan Digital Design 株式会社
toshiya.imoto@japan-d2.com

概要

意味的テキスト類似度 (Semantic Text Similarity) は，予め定めた基準に沿って２文書の類似度をスコア化するものである．金融分野ではニューステキストへの応用が考えられるが，地理的情報・エンティティ・時間・トピックなど，複数基準を総合判断する必要があるため，教師データセット作成の人的コストが高くなってしまう．本稿では，大規模言語モデル（LLM）を用い，少数のアノテーション済みサンプルを利用したニュース類似度の算出を試みた．特に，Few-shot、手がかり・推論の中間生成手法，自己整合性といったプロンプトエンジニアリング手法の有効性の検証を行い，SOTA の教師あり学習手法に迫る精度を達成した．更なる精度向上を目指すべく、地理的情報など各基準の類似度から段階的に解くサブタスク推論アプローチを検証し，提案モデルは SOTA と同水準の精度を達成した．

1 はじめに

意味的テキスト類似度 (Semantic Text Similarity) は，予め定めた基準に沿って２文書の類似度をスコア化するものである [1][2][3][4]．金融分野での主要なテキストの１つにニュースデータがある．ニュース分析はセンチメント分析 [5]，クラスタリング分析 [6]，金融資産との相関分析 [7]，ニュース推薦 [8] など多岐に渡る．ニュース類似度分析は，異なる２つのニュース記事について複数基準で類似度を測り，総合評価値を与えるタスクである [9]．クラスタリング分析やニュース推薦など，様々な用途に応用可能である．[9] では，各ニュース記事ペアについて複数人のアノテーターが，複数個の基準での類似度と総合類似度を付与したデータセットを提供している．ニュース記事は文章長が比較的長い事，地理的情報・エンティティ・時間・トピックなどの類似度を各々評価する事，各評価を組み合わせて総合評価を行う必要がある事など，データセット作成コストが高い事が課題である．

本稿では，ここ１，２年で急速に発展・応用が進む大規模言語モデル（LLM）を用い，少数のアノテーション済みサンプルを利用してニュース類似度の算出をし，高精度の分類モデルが構築可能か検証した．低資源下でのモデル構築を目指すため，学習データが必要な Fine-Tune は行わず，プロンプトエンジニアリングのみを利用した．実験データセット [9] は多言語ニュースから構成され，各ニュースの文脈を高度に理解する必要があるため，GPT-3.5 Turbo および GPT-4 Turbo (Azure OpenAI) を利用した．

精度の高いモデル構築を目指して，まず，入力テキストの工夫，CARP をベースにしたプロンプトエンジニア手法を実施した．結果，比較対象としたコンペティション１位の HFL モデル [10](本稿で SOTA モデルとも呼ぶ) に迫る高精度を達成した．更なる精度向上を目指すべく，地理的情報など各基準の類似度から段階的に解くアプローチ（本稿でサブタスク推論と呼ぶ）を検証し，提案モデルは SOTA モデルと同水準の精度を達成した．

2 予備知識

2.1 データセットとタスク

本稿では，SemEval 2022 Shared task8[11] を実験データセットとして利用する．[1)] 英語など 10 言語について，2020 年１月から６月までのニュース記事ペアを収集し，地理的情報 (GEO)，エンティティ（ENT），時間（TIME），ナラティブ（NAR），総合評価（OVERALL）の類似度を付与している[2)]．

- GEO: 地理的な近さ (場所、都市、国など)
- ENT: GEO で勘案した位置情報を除いた固有名詞 (人、会社、組織、商品、何かしらの生物)

1) アノテーションスキームやサンプルは [12] に纏まっている．

2) 実際は，記述スタイル（STYLE）とトーン（TONE）もあるが，総合評価に勘案されないため，本稿分析対象から割愛した

表 1 検証データのニュース記事ペアの数，平均文章数と平均トークン数

言語ペア	サンプル数	平均文章数	平均トークン数
it-it	401	8.42	581
de-de	272	12.42	624
ru-ru	194	8.54	805
zh-zh	161	1.05	818
de-en	115	10.88	530
es-en	105	7.51	589
en-en	103	19.33	768
es-es	67	3.75	510
zh-en	57	6.51	1354
tr-tr	50	8.65	677
ar-ar	39	2.17	830
pl-en	10	13.5	1000
pl-pl	10	14.05	778
fr-fr	1	4.00	347
-	1585	8.63	729

- TIME: 記事、あるいは、記事の内容の時期的な近さ
- NAR: 物語のスキームの近さ
- OVERALL: 実質的に同じニュースストーリーを網羅しているか？ (スタイル、トーンを除く)

各類似度は４種類のスコア，**VS(非常に似ている)，SS(いくらか似ている)，SD(いくらか似ていない)，VD(非常に似ていない)** の何れかを取る．なお，複数のアノテーターの平均値を用いるため正解ラベルは小数になり得る．表 1 に，検証データに含まれる言語ごとのニュース記事ペアの数，平均文章数，平均トークン数を載せた[3]．

コンペティションに倣い **OVERALL** の予測タスクを実施し，**VS,SS,SD,VD を各々 1,2,3,4 の数値に読み替え**，検証データセットでの**ピアソン相関係数による評価検証**を行った．

2.2 比較モデル（HFL）

提案モデルとの比較に，[11] で１位を獲得した HFL モデル [10] を採用した．xlm-roberta-large[13] ベースのアーキテクチャーで，データ前処理に２つの特徴，データ水増し（Data Augmentation）と Head-Tail 入力がある．データ水増しは，利用せずとも論文精度相当の結果を再現できたため本研究では実施していない[4]．Head-Tail 入力は，モデルの入力トークン数制限のため全文を入力せず，冒頭 200, 末尾 56 トークンを結合したテキストをモデルに入力する方法である．

2.3 CARP

CARP[14] は，LLM 自身に手がかり（CLUE）と推論（REASONING）を生成させ，続けて主題の分類タスクを解かせるというプロンプトエンジニア手法である．自己整合性（Self-Consistency）[15] は，同一クエリーに対し LLM に複数個の返答を出力させ，その平均値や最頻値を最終的な返答とする手法である．[14] では，16 個の手がかり・推論付き例示に加え自己整合性も利用し，既存手法の SOTA に匹敵する精度を達成した．

2.4 サブタスク推論

[16] では，タスクを解く際関連する子タスクを同時に解くことを考えた．シングルタスク推論（STI）とマルチタスク推論（MTI）の２つのアプローチで比較し，MTI の方が高精度であることが示されている．ここで，MTI は，マルチタスク学習同様，全ての子タスクと親タスクを LLM で一度に全て生成する手法を指す．STI は，各子タスクの返答生成を都度行い，親タスクの返答生成時には各子タスクの返答をプロンプトに追加して生成させる手法である．更なる精度向上のため CARP に加え，本稿で**サブタスク推論**アプローチと呼ぶ，GEO,ENT,NAR の各類似度（子タスク）を先に予測させ，続けて OVERALL（親タスク）を予測する手法を検証した．

3 実験計画

より高精度なニュース類似度算出モデルを模索するための実験を計画する．

実験 A1．入力トークン数の削減手法の比較
GPT-3.5 Turbo, GPT-4 Turbo モデルのトークン数上限は大きいものの，Few-shot の分ニュース記事ペアをプロンプトに書き込む必要があり，計算コストも高くなるため，トークン数は極力減らしたい．そこでトークン数削減方法として，Head-Tail 入力 (冒頭 200, 末尾 56 トークン) と要約入力（200 語以内）を比較する．

実験 A2．英語とマイナー言語の性能差の検証
[17] では，GPT-3.5, GPT-4 モデルでの低リソース言語（事前学習に占める割合が低い言語）の性能が英語などメジャー言語と比較し低いことが報告されている．そこで，非英語のニュース記事を英語に翻訳

3) 記事 URL のみ提供で本稿執筆時にアクセス不可な記事もあったため，取得可能な記事のみでデータセットを作成した．
4) 多言語セットアップでの Max Score が 81.8%（2 位は 80.1%）[9]，本研究での再現スコアが 80.69%(表 2).

表 2　全言語ペアでの Overall のピアソン相関

設定	モデル	入力テキスト	プロンプト	相関係数 (%)
1	GPT	Head-Tail	zero-shot	43.12
2	GPT	Head-Tail	3-shot	58.10
3	GPT	200 語要約	3-shot	55.88
4	GPT	Head-Tail (英語)	3-shot	54.50
5	GPT	Head-Tail	CARP 3-shot	63.75
6	GPT	Head-Tail	CARP 6-shot	64.98
7	GPT	Head-Tail	CARP 6-shot + SC	**75.25**
8	HFL	Head-Tail	-	80.69

して類似度を算出し精度比較する．

実験 A3．各種プロンプトエンジニアリング手法の有用性検証　Few-shot プロンプト，CARP，自己整合性プロンプト (SC) により，精度向上するか検証する．その結果，SOTA に匹敵するモデルは構築できたかを考察する．

実験 B．サブタスク推論のコンセプト正当性検証　実験 A で構築したモデルを更に発展させるため，GEO/ENT/NAR の各類似度を先に LLM に予測させ，続けて OVERALL を予測させるサブタスク推論アプローチを検証する[5]．**設定 6** をベースにした STI によるサブタスク推論アプローチを採用し，提案モデルが実験 A モデルより精度改善するか検証する．なお実験 B では，計算コスト削減のため en-en，es-en，de-en ペアのみを対象としている．また，GEO/ENT/NAR のいずれかに 4 より大きい数値が入ってる検証サンプルがあったため除外した[6]．

4　実験結果と考察

表 2 と表 3 に全実験結果を纏めた．実験設定の詳細は付録 A.1 に示す．

実験 A1　表 2 の**設定 2** と**設定 3** を比較し，少なくとも 3-shot プロンプトの場合では，記事要約を使うより冒頭・末尾を用いた方が高精度であることが分かった．これは [10] で議論されているように，重要な情報が記事冒頭・末尾に集中しているためと考えられる．Head-Tail 入力だと重要部分のみをモデルに入力するのに対して，要約だと重要でない部分も含めたニュース記事全体の情報が満遍なくモデルに入力され，精度が低くなると考えられる．

実験 A2　表 2 の**設定 2** と**設定 4** を比較し，少なくとも 3-shot プロンプトの場合では，低リソース言語の精度低下は確認できなかった．そこで本研究ではマルチリンガルのままモデル予測を行う事とした．

5)　TIME は，記事公表日などのメタデータをモデルで用いておらず予測に必要な情報が十分でないため除外した

6)　en-en は 68，es-en は 103，de-en は 94 サンプルとなった．

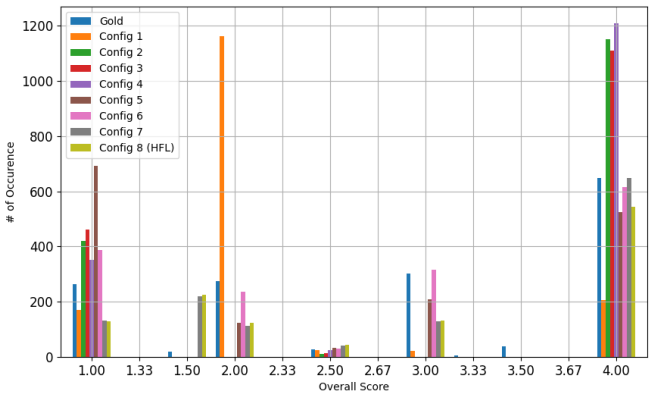

図 1　Overall の類似度の分布（Config は実験設定を表す）

実験 A3　表 2 の**設定 1** と**設定 2** を比較し，Zero-shot から Few-shot(3 例) にすることで，約 15%の精度向上を確認した．**設定 2** と**設定 5** を比較し，CARP（手がかり・推論の中間生成）で，5.7%の精度向上を確認した．**設定 6** と**設定 7** を比較し，SC(10 回の推論結果の平均値を用いた自己整合性プロンプト) により，10.3%の精度向上を確認した．また**設定 6** では，訓練データで誤回答の多かったタイプを追加で 3 例プロンプトに追加し，若干 (1.2%) の精度向上を確認した．更に，各実験設定での予測値分布を示す図 1 を見ると，CARP を導入する事で予測の偏りが大きく改善したことが分かる．ここまでの実験により，実験 A 最善モデル (**設定 7**) は 6 サンプルのみで 75.25%を達成し，HFL モデルの 80.69%に約 5%まで迫る結果を得られた．

実験 B　表 3 の**設定 9** と**設定 10** を比較し，サブタスク推論により全言語ペアで精度向上が見られ，全体では約 4%の精度向上を確認した．表 4 には各サブタスクの予測精度の一覧を載せている．**設定 10** と**設定 11** を比較し，**提案手法（STI）を用いる事で，HFL とほぼ同水準の全体精度を達成する**ことを確認した．

実験 B - 考察　**設定 10** の結果について，GEO/ENT/NAR（サブタスク）と OVERALL のスコアの関係性を考察したい．表 5 の**番号 1** と**番号 2** は，サブタスクの正解スコアについて各々平均値・最大値を取ったものと，OVERALL 正解スコアの相関係数を表す．ここから，OVERALL はサブタスクの最大値とより強く相関を持ち，特に de-en は，平均ではなく最大値を使う事で約 14%精度向上することが分かる．これは，サブタスクにどれか 1 つでも悪いスコアがあると，引っ張られて OVERALL スコ

表 3　3 言語ペアでの Overall のピアソン相関

			相関係数 (%)			
設定	モデル	サブタスク	en-en	es-en	de-en	3 ペア全体
9	GPT	-	77.46	86.73	72.75	80.85
10	GPT	STI	82.34	**87.70**	**79.49**	**84.72**
11	HFL	-	**83.23**	85.47	75.73	84.54

表 4　GEO・ENT・NAR の予測精度

	相関係数 (%)			
-	en-en	es-en	de-en	3 ペア全体
GEO	76.35	75.23	73.70	75.17
ENT	77.90	82.99	73.76	78.94
NAR	75.03	82.05	61.31	75.88

アも悪化する傾向があることを意味する．一方**番号 3** と**番号 4** は，正解の代わりに**設定 10** で算出したサブタスク・OVERALL 予測ラベル間の相関を取ったものである．ここから，モデルはサブタスクの最大よりも平均値に近い予測を行う事がわかる．つまり，モデルが最大値をより重視する予測ができれば，より高精度なモデルとなり得ることが示唆される．

STI モデルエラー分析　表 6 は，STI モデルに関する予測誤差（0.5 刻みで範囲を区切り集計）毎の件数の一覧である．ここでは，絶対値誤差が 2 より大きい en-en の一例に絞りエラー分析を実施する．付録 A.2 にモデルに入力した 2 記事の本文を，表 7 に各サブタスクと OVERALL の正解ラベルと予測を纏めた．まず後者から，ENT に一番大きな乖離があることが分かる．そこで本文を比較すると，共通したエンティティはニュースの主題と関係が薄いものが多い事に気づく（「居場所確保命令」は記事 2 の主題ではない，「エイミー・グラフ」は記者の名前，など）．ここで，Head-Tail 入力を採用しているため記事の中間部分は割愛されてしまう事に注意したい．

実際記事 2 の原文は，コロナウィルスの検査能力の拡大が主題で，PCR 技術，検査に必要な物質の供給不足，支援者情報などに関して，記事 1 にない多くのエンティティが記載されていた．Head-Tail に書かれた主題と関係が薄いエンティティの共通性をモデルが重視してしまい誤分類（ENT の類似度が過度に高くなった）したものと考えられる．

5　おわりに

本稿では，LLM を用い少数のアノテーション済みサンプルを利用したニュース類似度の算出を試みた．

表 5　3 言語ペアでの GEO/ENT/NAR と OVERALL のピアソン相関

			相関係数 (%)			
番号	サブタスク	OVERALL	en-en	es-en	de-en	3 ペア全体
1	正解（平均）	正解	85.60	90.46	76.23	86.67
2	正解（最大）	正解	**87.20**	**90.99**	**90.55**	**90.66**
3	予測（平均）	予測	**96.08**	**96.88**	**92.03**	**95.58**
4	予測（最大）	予測	93.48	93.43	89.15	93.15

表 6　STI モデルの絶対予測誤差（Err）ごとの件数

Err	en-en	es-en	de-en
Err ≤ 0.5	44	72	79
0.5 < Err ≤ 1	19	26	11
1 < Err ≤ 1.5	2	3	1
1.5 < Err ≤ 2	2	2	3
2 < Err	1	0	0
-	68	103	94

表 7　誤分類サンプルの詳細

-	正解ラベル	予測
GEO	1.5	1.0
ENT	3.5	1.4
NAR	3.0	1.67
OVERALL	3.5	1.2

Few-shot、CARP，自己整合性といったプロンプトエンジニアリング手法に加え，GEO/ENT/NAR の各基準の類似度から段階的に解くサブタスク推論アプローチを提案し，SOTA と同水準の精度を達成した．なお，本稿では MTI を実施しなかった．入力長が長くなりコストが高くなる事，ENT など固有表現抽出に関しては新しいエンティティの追加もあり更新頻度の低い GPT モデルより専用のモデルを用意した方がよい事などが主な理由であるが，他タスクへの適用も見据え有効性の追加検証を実施したい．今後は他にも，検証サンプル数の増加，Head-Tail 入力トークン数の増加，CARP の例示数の増加，記事日付等を用いた TIME の予測，GPT モデルの学習期間外での精度検証などを行い，更なる精度向上を目指したい．

参考文献

[1] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. Association for Computational Linguistics, 2017.

[2] Akshita Jha, Vineeth Rakesh, Jaideep Chandrashekar, Adithya Samavedhi, and Chandan K. Reddy. Supervised contrastive learning for interpretable long-form document matching. Vol. 17, No. 2, 2023.

[3] Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. Sub-sentence encoder: Contrastive learning of propositional semantic representations, 2023.

[4] Shuhe Wang, Beiming Cao, Shengyu Zhang, Xiaoya Li, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Simgpt: Text similarity via gpt annotated data, 2023.

[5] Sorouralsadat Fatemi and Yuheng Hu. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis, 2023.

[6] Kailash Karthik Saravanakumar, Miguel Ballesteros, Muthu Kumar Chandrasekaran, and Kathleen McKeown. Event-driven news stream clustering using entity-aware contextual embeddings. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2330–2340, Online, April 2021. Association for Computational Linguistics.

[7] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models, 2023.

[8] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. Exploring fine-tuning chatgpt for news recommendation, 2023.

[9] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. SemEval-2022 task 8: Multilingual news article similarity. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1094–1106, Seattle, United States, July 2022. Association for Computational Linguistics.

[10] Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1114–1120, Seattle, United States, July 2022. Association for Computational Linguistics.

[11] Semeval 2022 task 8: Multilingual news article similarity, 2022. https://competitions.codalab.org/competitions/33835.

[12] Semeval 2022 task 8: Multilingual news article similarity dataset details, 2022. https://zenodo.org/records/6507872.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

[14] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models, 2023.

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

[16] Multi-task inference: Can large language models follow multiple instructions at once?, 2023. https://openreview.net/forum?id=_HP30A8V3DT.

[17] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

A Appendix

A.1 実験設定の詳細

設定 1 から 4 図 2 のプロンプトを用い，温度は 0 で GPT-3.5 Turbo(gpt-35-turbo-16k-0613) により生成した．[12] から VS/SS/SD（VD の例はなかった）を 1 例ずつ選出し 3-shot に用いた．**設定 3** の 200 語要約は LLaMA-13B-Chat (https://huggingface.co/meta-llama/Llama-2-13b-chat-hf) にて，簡易なプロンプトを使って行った．**設定 4** の英語への翻訳は https://github.com/ssut/py-googletrans にて行った．

設定 5,6 図 3 のプロンプトを用い，温度は 0 で GPT-3.5 Turbo により生成した．全体の数%は予測がうまくいかなかったため，それらは GPT-4 Turbo (gpt-4-1106-preview) で再生成した．**設定 6** で追加した 3-shot は，訓練データ全体を**設定 5** で推論しモデルが VS と誤回答するケースが多い事が判明したため，正答が SS,SD,VD のケースを各 1 例選出し，CLUES と REASONING を人手で作成し追加したものである．

設定 7 温度を 0.7 に設定し 10 個の返答と類似度を生成，その平均値を最終的な類似度とした．

設定 9 **設定 7** の 3 言語ペアに絞った結果を載せた．

設定 10 図 4 のプロンプトを用い，**設定 7** と同様 10 個の平均値を算出した．GEO,ENT,NAR の各サブタスク予測は，CARP+SC(**設定 7**) をベースに，Instruction や例を各々カスタマイズしたものを用いた．

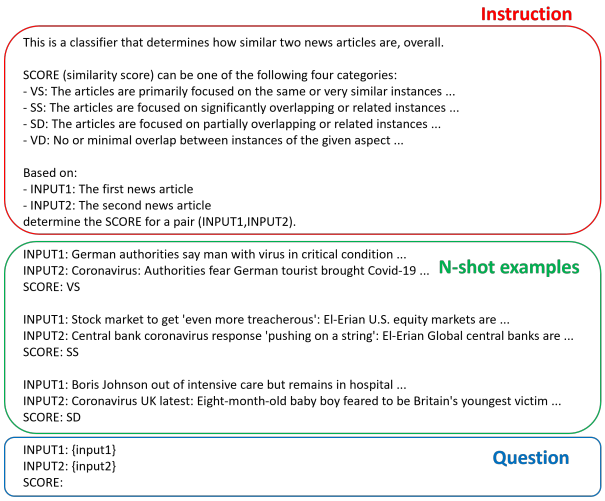

図 2 few-shot プロンプト

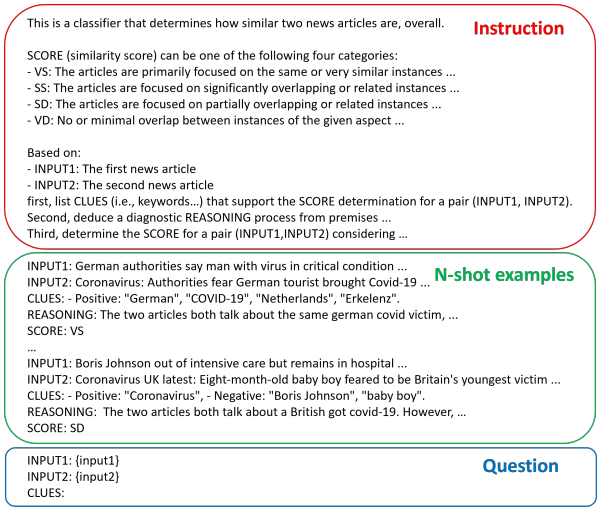

図 3 carp プロンプト

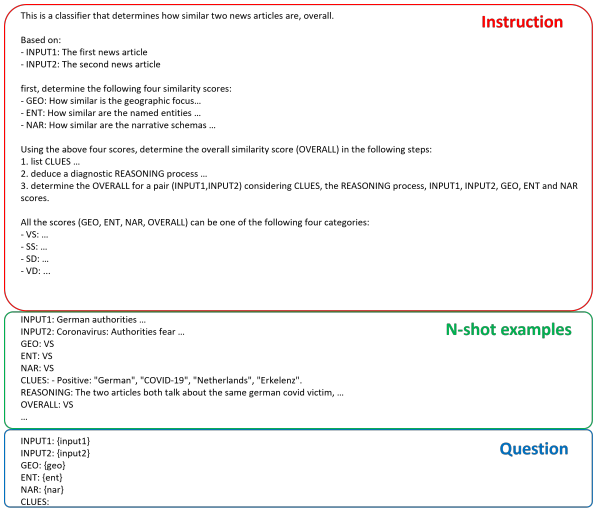

図 4 サブタスク推論プロンプト

A.2 エラー分析の記事詳細

記事 1
コロナウイルス更新：サンフランシスコの症例数が105に急増： 2020年3月20日、カリフォルニア州サンフランシスコのゴールデンゲートブリッジにはほとんど観光客が訪れていません。ベイエリアは、COVID-19コロナウイルスのため、居場所を確保する命令の下にあります... 写真：ダグラス・ジマーマン/SFGate 画像1 / 60 キャプション閉じる ... こちらでコロナウイルスに関する「The Daily」ニュースレターにサインアップしてください。エイミー・グラフはSFGateのデジタルエディターです。

記事 2
コロナウイルス更新：UCSF、検査能力を1日あたり100人から1,000人に拡大 - ギャラリーをスクロールして、COVID-19の拡散を防ぐための居場所確保命令下にあるサンフランシスコ・ベイエリアの画像をご覧ください。カリフォルニア大学サンフランシスコ校は木曜日に、今後数週間でCOVID-19検査能力を増加させることを目指していると発表しました。UCSFヘルスは、疾病管理予防センターが行っているものと似た「ポリメラーゼ連鎖反応（PCR）技術」を使用したテストで、...を検査しています。UCSFは声明で、「PCRにより、...」と述べています。この技術はグラフに依存しています。グラフはSFGATEのデジタルエディターです。彼女にメールしてください...

図 5 日本語に翻訳した 2 記事の本文（一部省略）