# 自動ファクトチェックのための事実の分解による含意関係認識

雨宮正弥 [1]　狩野芳伸 [1]
[1] 静岡大学大学院 総合科学技術研究科 情報学専攻
{mamemiya,kano}@kanolab.net

## 概要

含意関係認識は、ファクトチェック自動化において有望なアプローチの一つである。しかし、既存の含意関係認識手法をそのまま適用するだけでは、正確な検証が難しい場合がある。本研究では、前提文や仮定文を個別の事実に分解し、分解した要素に対して含意関係認識を行う手法を提案する。独自に構築したファクトチェック性能評価データセットにおいて、提案手法を導入することで LLM 単独よりも gpt-4o において 9.34、llmjp-3 において 20.67 ポイントの Accuracy 向上を達成し、その有効性を示した。

## 1　はじめに

インターネットや SNS の普及に伴い、情報が迅速に拡散される一方で、不正確な情報や誤情報の拡散が問題となっている。そのため正確な情報を確認し、不正確な情報を訂正するファクトチェックの重要性は高い。しかし、人力によるファクトチェックは、膨大な時間と労力を要するため、自動化が求められている。その中で、含意関係認識は、2 つの文間の論理的関係を判定する技術として、自動ファクトチェックシステムにおいて有望な手法の一つである。本稿では自動ファクトチェックの基盤技術として、文ペアの含意関係認識を行う。

含意関係認識とは、前提文と仮定文のペアについて、前提文が真であると仮定した際に仮定文が真であると判断できれば「含意」、仮定文が偽であると判断できれば「矛盾」、どちらとも判断できない場合は「中立」と分類するタスクである。しかし、この技術をファクトチェックに適用する際には、仮定文全体が前提文に含意されない限り「含意」と判定されないという課題がある。例えば、「今日の試合で負けたなんて信じられない」という仮定文がある場合、「今日の試合で負けた」という部分が前提文に含意されていたとしても、「信じられない」という部分が原因で、仮定文全体としては「含意」と判断されない可能性がある。

我々は、前提文や仮定文を事実に分解し、分解された要素に対して含意関係認識を行う手法を提案する。独自に構築したファクトチェック評価データセットにおいて、提案手法の導入により LLM 単独よりも **gpt-4o** において 9.34、**llmjp** において 20.67 ポイントの Accuracy 向上を達成し、有効性を示した。

## 2　関連研究

### 2.1　含意関係認識

日本語の含意関係認識データセットには、JSNLI[1]、JSICK[2]、JNLI[3] 等がある。それぞれ事前学習済み BERT[4] をファインチューンした Accuracy の報告値をカッコ内に表記して紹介する。

JSNLI[1] は英語の大規模な含意関係認識データセットである SNLI[5] を日本語に機械翻訳したものである（0.929）。JSICK[2] は様々な言語現象を含む英語の含意関係認識データセットである SICK[6] を人手で日本語に翻訳したものである（0.84）。JNLI[3] は、日本語理解ベンチマーク JGLUE[3] に含まれる含意関係認識データセットで、翻訳を介さずに作成された（0.906）。JNLI はクラウドソーシングで構築され、画像の内容を文章で表現させることで、含意と中立のラベルを持つ文ペアを作成している。矛盾ラベルの文ペアは、表現した文章に対して矛盾する内容をクラウドソーシングで構築された。

OpenAI 社の GPT-4o[7] は、さまざまなベンチマークにおいて人間に匹敵するパフォーマンスを示している。Nejumi LLM リーダーボード 3[8] で公開されたデータによると、gpt-4o-2024-11-20 を用いて JNLI や JSICK といった意味解析タスクを評価した結果、Accuracy 0.795 を達成している。

また、NLI タスクにおいて高性能な日本語対応のオープンモデルとして、LLM-JP が挙げられる。Nejumi LLM リーダーボード Neo[9] によれば、JNLI の性能が最も高いモデルは llm-jp-13b-instruct-full-

jaster-v1.0[1)]であり、Accuracy は 0.91 を達成している。ただし、高い性能を示している理由として、インストラクション訓練データに JNLI を含んでいることが挙げられる。

### 2.2 含意関係認識を用いたファクトチェック

栗原ら [10] は、疑義言説を仮定文、関連文書を前提文として、JSNLI データセットが含まれる FCSNLI データセットを用いて BERT をファインチューニングし、3 値分類において F 値 0.538 を達成した。

我藤ら [11] は、議事録の要約を仮定文、議事録本文を前提文として扱い、タスク用に作成された訓練データセットでファインチューニングした BERT モデルを用いて含意関係認識によるファクトチェックを実施し、F 値 0.9183 を達成した。

英語においても含意関係認識を用いたファクトチェックの研究が進められている。[12][13]

## 3 提案手法

従来の含意関係認識手法をファクトチェックにそのまま適用するだけでは、正確な検証が難しい場合がある。たとえば、日本ファクトチェックセンターの検証記事[2)]において、「遺族年金廃止とは頭おかしいよ…（後略）」という検証対象文が取り上げられ、「遺族年金廃止」に焦点を当てたファクトチェックの結果、「誤り」と判定された。この場合、前提文「遺族年金の廃止が考えられていないと発表された」と仮定文「遺族年金廃止とは頭おかしいよ…」を従来の含意関係認識手法で解析すると、結果は「中立」と判断され、適切な検証が行えない。

本研究では、前提文や仮定文を個別の「事実」に分解し、分解したそれぞれの事実に対して含意関係認識を適用することで、この課題を解決することを目指す。ここでいう「事実」とは、述語が一つだけ含まれる粒度のフレーズに対応する。たとえば、検証対象文「遺族年金廃止とは頭おかしいよ…」を「遺族年金を廃止する」と「遺族年金廃止とは頭おかしいよ…」の 2 つのフレーズに分解する。このうち「遺族年金廃止」というフレーズは、前提文「遺族年金の廃止が考えられていない」に矛盾するため、全体として検証対象文が誤りであると判定できる。

我々の提案手法では、前提文と仮定文を分解した後、それぞれの分解フレーズについて総当たりで含意関係認識を実行し、最終的に仮定文全体として「含意」「中立」「矛盾」のいずれかにラベル付けを行う (図 1)。本研究における各ラベルの定義は以下の通りである。

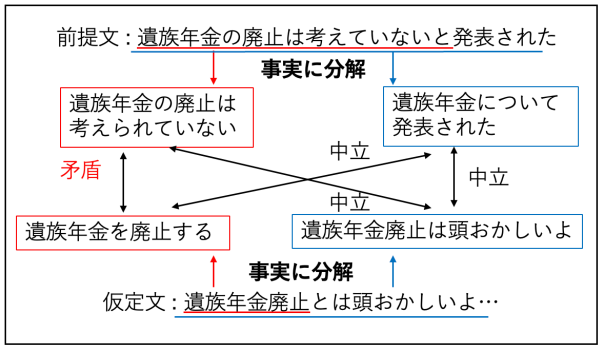


**図 1** 提案手法の概要

**含意の定義** 仮定文の内容が前提文に対して含意している部分があり、それ以外の部分が中立である

**矛盾の定義** 仮定文の中に前提文に対して矛盾している部分がある

**中立の定義** 上 2 つ以外の場合（仮定文の内容が前提文に全て中立）

### 3.1 事実分解

ファクトチェックは、SNS 投稿を対象に行うことが多い。しかし、SNS 投稿のテキストは形式的な文法に従わないケースが多く、既存のルールベースの文解析手法の適用が難しい。このため、本研究では LLM（Large Language Model、大規模言語モデル）を用いた事実分解の方法を採用した。使用する LLM には、表現力、要約能力、抽出能力といった汎用的な言語性能が求められる。そこで、Nejumi LLM リーダーボード 3 で汎用言語性能の平均評価値が優れており、特に抽出タスクで高精度を示している gpt-4o-2024-11-20 を使用する。

文章中に含まれる複数の事実を分解し個別に抽出することを目的とした図 2 のプロンプトを設計した。このプロンプトでは、指示語の使用を禁じ、主語や目的語が明確に出力されるよう工夫している。

### 3.2 含意関係認識

含意関係認識を行う LLM として、JNLI 性能が高い llm-jp-13b-instruct-full-jaster-v1.0 と、Nejumi LLM リーダーボード 3 の意味解析タスクで高い性能を示し、文分解にも使用した gpt-4o-2024-11-20 を使用

1) https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-v1.0
2) https://www.factcheckcenter.jp/fact-check/lifestyle/false-end-of-survivor-pension/

```
###指示###
文章で示唆されている事実を列挙してください。
出力に「この、その」等の指示語を使わないでください。

###文章###
{text}

###出力形式###
[事実 1]*[事実 2]*[事実 3]
```

**図 2**　事実の分解を指示するプロンプトテンプレート

する。

プロンプトについては、両モデルともに、LLM-jp 評価スクリプトで JNLI タスクに使用されていたもの[3]を使用した（図 5）。

# 4　実験

図 5 の含意関係認識プロンプトのみでファクトチェックをする手法と、図 2 に示した事実分解のプロンプトを使って事実を抽出してからファクトチェックをする手法の性能評価と比較を行った。

## 4.1　評価用データセットの構築

日本ファクトチェックセンターの記事[4]と、X[5]から収集したデータを基に、独自のファクトチェック性能評価用のデータセットを構築した[6]。

日本ファクトチェックセンターの記事では、対象となる言説のほとんどに「誤り」または「不正確」のラベルが付与されている。そのため、本研究では主に「矛盾」ラベルを付与するデータとして活用した。記事は「検証対象」「検証過程」「判定」で構成されている。「検証対象」は、主に X のポストが対象となっており、それを前提文として使用した。「検証過程」は、検証に使用した情報や根拠が記載されているため、それを仮定文として使用した。

主に「含意」「中立」ラベルを付与するデータとして、X の「Yahoo!ニュース」公式アカウント（@YahooNewsTopics）によるニュース記事ポストと、それに対するリプライを収集した。ニュース記事本文を前提文、リプライを仮定文として使用した。

これらの前提文と仮定文のペアに対し、3 人のアノテーターが前節で記述した基準に基づいて「含意」「中立」「矛盾」の 3 値でラベル付けを行った。そのうち 2 人以上が同じラベルをつけたものを採用し、それ以外は破棄した。各クラスのデータを 50 件ずつ揃え、計 150 データで評価を行った。

## 4.2　手法

以下の 4 つの手法を比較し、ファクトチェックの性能を評価した。いずれも LLM を用いゼロショットで含意関係認識を行う。提案手法は前提文と仮定文を分解してから総当たりで含意関係認識を実行させる。

| 手法名 | 種別 | LLM モデル名 |
|---|---|---|
| **gpt-4o** | ベースライン | gpt-4o-2024-11-20 |
| **分解+gpt-4o** | 提案手法 | gpt-4o-2024-11-20 |
| **llmjp** | ベースライン | llm-jp-13b-instruct-full-jaster-v1.0 |
| **分解+llmjp** | 提案手法 | llm-jp-13b-instruct-full-jaster-v1.0 |

**表 1**　実験に用いた手法の一覧

# 5　結果

前節で示した 4 通りの手法について評価した結果を表 2 に示す。最も高い性能（accuracy）を示したのは**分解+gpt-4o** であった。この手法は、precision, recall, F1 スコアにおいても他の手法を上回った。一方、分解を行わずに含意関係認識を行った **gpt-4o** や **llmjp** では、特に「含意」における評価値が著しく低く、Recall はそれぞれ 0.12 と 0.10 に留まった。前提文や仮定文を分解する提案手法が LLM の種類を問わず大きく貢献したといえる。

**分解+llmjp** は**分解+gpt-4o** と比較して全ての評価指標が低くなった。今回使用した llm-jp-13b-instruct-full-jaster は JNLI でファインチューンされており、過学習して一般性能が低下した可能性がある。

**分解+gpt-4o** は、「矛盾」の評価値が **gpt-4o** を下回ったが、他の評価指標での優位性を考慮すると、提案手法と高性能な LLM を組み合わせた場合が全体的に最も有効であったといえる。

# 6　分析

## 6.1　提案手法によって正解できるようになった事例

提案手法を用いることで、従来手法では見逃されていた含意関係を把握できるようになった事例を図 3 に示す。なお、本稿の事例は匿名化のため、いずれも実投稿を参考に同等の文構造をとるよう作例し

3）https://github.com/llm-jp/llm-jp-eval/blob/dev/src/llm_jp_eval/jaster/jnli.py
4）https://www.factcheckcenter.jp/
5）https://x.com/
6）無償で一般に公開予定である

| モデル名 | Acc | 矛盾 (50 件) | | | | 含意 (50 件) | | | | 中立 (50 件) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 正解 | P | R | F1 | 正解 | P | R | F1 | 正解 | P | R | F1 |
| **gpt-4o** | 61.33 | 37 | 94.87 | 74.00 | **83.15** | 6 | 100.00 | 12.00 | 21.43 | 49 | 46.67 | 98.00 | 63.23 |
| **llmjp** | 38.00 | 8 | 53.33 | 16.00 | 24.62 | 5 | 100.00 | 10.00 | 18.18 | 44 | 34.38 | 88.00 | 49.44 |
| **分解+gpt-4o** | **70.67** | 38 | 67.86 | 76.00 | 71.70 | 35 | 71.43 | 70.00 | **70.71** | 33 | 73.33 | 66.00 | **69.47** |
| **分解+llmjp** | 58.67 | 35 | 48.61 | 70.00 | 57.38 | 22 | 66.67 | 44.00 | 53.01 | 28 | 62.22 | 56.00 | 58.95 |

**表 2**　実験結果 (Acc: Accuracy(%), P: Precision(%), R: Recall(%), F1: F1-Score(%)

たものである。図 3 の事例 1-3 では、従来手法ではすべて「中立」と判定したが、**分解+gpt-4o** では正確な判定が可能となった。

事例 1 では、前提文を分解して得られた「講演会が中止になることが発表された」と仮定文を分解して得られた「講演会が中止になった」を比較し、「含意」と判定できた。従来手法では「信じられない」のような感想部分が影響し中立と判断されていたが、提案手法では仮定文中の事実部分を抽出することで正確な検出が可能となった。

事例 2 では、前提文を分解して得られた「講演会が中止になることが発表された」と仮定文を分解して得られた「お祭りが中止である可能性がある」を比較し、「含意」と判定できた。従来手法では仮定文内の並列要素（講演会とお祭り）を一括処理し、中立と判断していたが、提案手法ではこれを分解して個別に検証することで正確な判定が可能となった。

事例 3 では、前提文を分解して得られた「山田さんが芸能界を引退する」と仮定文を分解して得られた「山田さんがサッカー界を引退する」を比較し、「矛盾」と判定できた。従来手法では複数の事実が混在する文中に矛盾要素が含まる場合、矛盾を検出できないことがあるが、提案手法では文を分解して検証することで矛盾を正確に検出できた。

**事例 1**
前提文 文化ホールで予定されていた有名人の講演会が中止になることが発表された。
仮定文 講演会が中止になるなんて信じられない

**事例 2**
前提文 文化ホールで予定されていた有名人の講演会が中止になることが発表された。
仮定文 講演会もお祭りも中止か…

**事例 3**
前提文 山田さんが芸能界を引退することを発表した。
仮定文 山田さんは不祥事起こしちゃったからサッカー界を引退するんだね

**図 3**　提案手法によって判定できるようになった事例

## 6.2　提案手法でも失敗する事例

失敗の原因は「部分的な情報に基づく誤判定」と「時系列情報の欠如」が主であった。

**分解+gpt-4o** で判定に失敗した事例を図 4 に示す。

失敗事例 1 では、前提文を分解して得られた「NHK は広告放送を禁じられている」と仮定文を分解して得られた「番組が広告付きで配信されている」を比較し、「矛盾」と判定した。全体の文脈を考慮すれば矛盾ではないが、部分的な情報に基づいて矛盾と誤判定してしまった。分解された情報間の意味的な文脈や関係を適切に考慮する必要がある。

失敗事例 2 では、前提文を分解して得た「10 年ぶりに免許を取得した」と、仮定文を分解して得た「運転免許を取得した」を比較し「含意」と判定した。しかし時系列を考慮すると、全体としては矛盾である。この失敗は、現在の分解手法が時系列情報を十分に考慮できていないためと考えられる。

**失敗事例 1**
前提文 NHK は、広告放送を禁じられているため、営利目的の放送との誤認を避けるために、番組の配信停止を要求した。
仮定文 何者かが番組を広告付きで配信していることに対して NHK が抗議している。

**失敗事例 22**
前提文 今日、10 年ぶりに自動車運転免許を取得した。10 年前、飲酒運転をした。
仮定文 10 年ぶりの運転免許取得の後にコレだ。飲酒運転

**図 4**　提案手法が判定に失敗した事例

# 7　おわりに

前提文や仮定文を個別の事実に分解し、分解された要素に対して含意関係認識を行う手法を提案した。独自に作成したファクトチェック性能評価用データセットを用いた実験では、LLM と組み合わせることで性能の向上を示した。今後は、分解した要素間の文脈や関係性を適切に考慮する仕組みを導入し、さらなる性能向上を目指す。

## 謝辞

本研究はJSPS科研費（JP22H00804）、JSTさきがけ（JPMJPR2461）、JST AIP加速課題（JPMJCR22U4）、およびセコム科学技術財団特定領域研究助成の支援をうけた。

## 参考文献

[1] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会 第 244 回自然言語処理研究会, Vol. 2020-NL-244, No. 6, pp. 1–8, 7 2020.

[2] 谷中瞳, 峯島宏次. Jsick: 日本語構成的推論・類似度データセットの構築. 人工知能学会全国大会論文集, Vol. JSAI2021, pp. 4J3GS6f02–4J3GS6f02, 2021.

[3] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[6] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[7] OpenAI : Aaron Hurst and et al. Gpt-4o system card, 2024. 2410.21276 https://arxiv.org/abs/2410.21276.

[8] Nejumi LLM リーダーボード 3,（2024-12 閲覧）. https://wandb.ai/wandb-japan/llm-leaderboard3/reports/Nejumi-LLM-3--Vmlldzo3OTg2NjM2.

[9] Nejumi LLM リーダーボード Neo,（2024-12 閲覧）. https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0.

[10] 健太郎栗原, 大輔河原. ファクトチェック支援のための含意関係認識システム. 言語処理学会 第 27 回年次大会 発表論文集, pp. 1734–1739, 03 2021.

[11] 勇樹我藤, 友良秋葉. パッセージ検索と含意関係認識による議会議事録を対象としたファクトチェック. 言語処理学会 第 28 回年次大会 発表論文集, pp. 768–772, 03 2022.

[12] Aalok Sathe and Joonsuk Park. Automatic fact-checking with document-level annotations using BERT and multiple instance learning. In Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, **Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)**, pp. 101–107, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. Automated fact-checking of claims from Wikipedia. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 6874–6882, Marseille, France, May 2020. European Language Resources Association.

# A 付録

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。
要求を適切に満たす応答を書きなさい。
### 指示: 前提と仮説の関係を `entailment`、`contradiction`、`neutral` の中から回答してください。
それ以外には何も含めないことを厳守してください。
制約：

- 前提から仮説が、論理的知識や常識的知識を用いて導出可能である場合は `entailment` と出力
- 前提と仮説が両立しえない場合は `contradiction` と出力
- そのいずれでもない場合は `neutral` と出力

### 入力:
**前提**：{zentei}
**仮定**：{katei}
### 応答:

**図 5** 含意関係認識のプロンプトテンプレート