統計的キーワード抽出と類推による人間の認知過程に基づく語義曖昧性解消

伊藤碧天 [1]　佐々木稔 [2]
[1] 茨城大学　工学部
[2] 茨城大学　理工学研究科　情報科学領域
{22t4006r,minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

語義曖昧性解消（WSD）において大規模言語モデルは高い精度を示すが，推論プロセスの不透明さが課題である．本論文では，重要語への着眼と類推という人間の認知プロセスを外部化した，解釈可能な WSD 手法を提案する．本稿では特に「どのキーワード抽出が文脈の固定に寄与するか」を検証するため，統計的・意味的手法の比較実験を行った．FEWS[1] を用いた実験の結果，統計的な BM25[2] による抽出が Zero-shot[1)] で 76.9%の正答率を記録し，既存の専用学習モデルである ESR[3]（75.8%）や LLM の CoT（56.0%）を上回る性能を達成した．本結果は，外部的な構造化ステップが精度向上と判定根拠の透明化を両立させる上で極めて有効であることを実証している．

1　はじめに

語義曖昧性解消（WSD）は，文脈に応じて多義語の適切な意味を選択する自然言語処理の基幹タスクである．近年，大規模言語モデル（LLM）の台頭により，Chain-of-Thought (CoT) [4] 等の内部推論を介した手法が高い精度を収めている．しかし，これらの手法は推論過程がモデル内部のパラメータに閉じており，なぜ特定の語義が選ばれたのかという判定根拠を人間が客観的に検証することは依然として困難である．また，文脈中のどの語が語義判断に寄与したのかが明示されないため，人間の語義判断プロセスとの対応関係が不明瞭である．

本研究では，このブラックボックス性を打破するために，情報検索に基づく外部アンカリングと LLM による意味類推を組み合わせた，新しい Zero-shot WSD 手法を提案する．提案手法は，(1) 文脈キーワード抽出，(2) 類推例文生成，(3) 多角的多数決判定，という 3 段階のパイプラインから構成される．

この構成により，本手法は次の特徴を有する．第一に，BM25 により抽出された重要語や生成された例文を通じて，語義選択の根拠を人間が事後的に追跡可能である点である．第二に，教師データを一切用いない Zero-shot 設定においても，既存の Zero-shot WSD 手法を上回る性能を達成できる点である．第三に，語義判断を単一の出力に依存せず，複数の類推結果を信頼度付きで統合することで，推論の安定性を高めている点である．本稿において特に焦点を当てるのは，第一段階であるキーワード抽出手法の有効性の検証である．

2　関連研究

2.1　知識ベースおよび語義定義に基づく WSD

語義曖昧性解消（WSD）の初期研究では，WordNet[5] 等の語彙資源や語義定義（gloss）を用いた知識ベース手法が広く用いられてきた．代表的なアプローチとして，Lesk[6] に代表される，文脈中の語と語義定義文との語彙的重なり（overlap）に基づいて語義を決定する手法が提案されている．これらの手法は教師データを必要とせず，語義判断の根拠を定義文や一致語として明示できるという利点を持つ一方，表層的な語彙一致に強く依存するため，語彙の言い換えや文脈的推論を十分に捉えられないという限界がある．

2.2　教師あり WSD

BERT などの事前学習言語モデルを用いた教師あり WSD 手法は，文脈表現と語義表現を統合的に学習することで高い性能を達成しているが [7]，大量

1）　Zero-shot とは LLM にタスク特化の教師あり学習を行っていない状態を指す

の教師データを必要とするという制約がある．これに対し，FEWS データセットの登場以降，限られた例文のみを用いる Few-shot および Zero-shot WSD が注目されている．既存研究では，語義定義や例文を埋め込み空間に写像し，文脈との類似度に基づいて語義を判定する手法が主流であるが，教師データを用いない完全な Zero-shot 設定においては，性能と推論の安定性を同時に確保することが依然として課題である．

2.3 LLM を用いた WSD

大規模言語モデル（LLM）は，明示的なタスク学習なしに多様な自然言語処理タスクを実行できる能力を示しており，WSD においても Zero-shot 推論が可能であることが報告されている．近年の研究では，語義定義をプロンプトとして与え，言語モデルに直接語義選択を行わせる手法や，Chain-of-Thought による推論過程を導入する試みがなされている．その中でも，特に GlossGPT[8] は GPT を CoT を用いて，多くのデータセットで SOTA を達成した．

しかし，これらの手法では，モデル内部の推論過程が外部から直接観測できず，語義選択の根拠が不透明であるという課題が指摘されている．また，単一の推論結果に依存するため，出力のばらつきや不安定性が性能低下につながる場合がある．

3 提案手法

本研究では, 人間が未知の多義語に遭遇した際に, 周囲の単語（手がかり）を特定し, 自身の知識から類似した使用事例を想起して判断を下すという認知プロセスに着目した. 提案手法は, このプロセスを以下の 3 段階のパイプラインとして構造化し, LLM に実行させるものである. 図 1 は提案手法のパイプラインを図にしたものである．

3.1 ステップ 1：統計的指標に基づく文脈の固定

本ステップの目的は，対象文 S の中から, 対象語 w の語義を決定付ける上で重要な手がかりとなるキーワード $K = \{k_1, k_2, \cdots, k_5\}$ を抽出し，推論の根拠を外部に明示することである．本研究では，この文脈の固定において，情報検索分野で広く用いられる統計的指標 BM25 [9] を採用する．この選択には，本論文の核となる「プロセスの透明化」と「解釈可能性」の観点から，以下の 2 つの重要な意図がある．

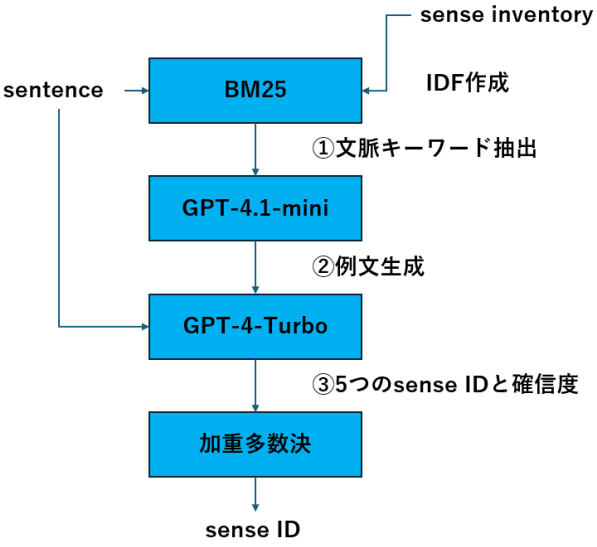

図 1 パイプライン図

1. **Attention のブラックボックス性への介入**:LLM の内部 Attention 機構は，文中の広範な依存関係を捉えられる一方で，どの単語が最終的な語義判定にどの程度寄与したかを人間が直感的に理解することは困難である．あえて古典的な BM25 を用いることで，単語頻度と希少性に基づいた「客観的かつ再現可能な証拠」を抽出プロセスに導入し，後続の LLM 推論が注目すべき対象を物理的・明示的に制限する．
2. **統計的アンカーと意味的アンカーの比較検証**: 本稿では，LLM 自身の能力に依存してキーワードを選ぶ「意味的抽出（LLM-Select）」に対し，統計的な客観性を持つ BM25 がどのように機能するかを検証対象とする．LLM 自身にキーワードを選ばせる手法は，モデルが既に持っている語義の偏りを助長する恐れがある．これに対し，BM25 によるキーワード抽出は，モデルの外部から統計的事実を突きつける役割を果たし，推論のハルシネーションを抑制する「思考の足場」として機能する．

このように，ステップ 1 でキーワードを露出させることは，システムが**文脈のどこに着目したか**を人間が事後的に検証することを可能にし，WSD における判定精度の向上と解釈可能性の確保を両立させる基盤となる．

3.2 ステップ 2：類似事例の動的生成

次に，ステップ 1 で特定された各キーワード k_i を核として, 対象語 w がその文脈でどのように使わ

れるかの具体例 (類推例文)E_i を GPT-4.1-mini を用いて生成する．このとき，GPT-4.1-mini で生成する例文は一つのキーワードにつき 1 文である．これは LLM がキーワード k_i と対象語 w の関係をどのように解釈しているかを自然言語で出力させる役割を持つ．単に対象文のみを参照するブラックボックスな推論とは異なり，LLM 自ら生成した「典型的な使用事例」を参照点として持つことで, 語義の境界線をより鮮明に認識することが可能となる. なお, この工程には, 生成能力とコスト効率のバランスに優れた GPT-4.1-mini を採用した.

3.3 ステップ 3: 加重多数決による判定

最後に, 最高水準の推論能力を持つ GPT-4-Turbo を用い, 最終的な語義を決定する．ここでは，対象語 w の語義候補リスト D と, キーワード k_i および生成された例文 E_i のセットを順次入力し，最も整合性の高い語義の組み合わせを探索させる.

本ステップの特徴は, 5 つの独立したキーワードに基づくアンサンブル判定にある. 各推論パスにおいて，モデルには「High（確信）」,「Medium（中程度）」,「Low（疑義）」の 3 段階の確信度を付与させる. 最終判定では，これらの確信度を数値化（High=1, Medium=0.4, Low=0.1）して加重集計を行う．確信度の数値化については，複数回の検証により決定した．具体的には，それぞれの段階を重く見るため，High の選択 1 つが Medium の選択 2 つに決定を覆されることのないようにこの値とした．Low に関しては，ほとんど加重を掛けておらず，Low のみの選択肢だった場合に，0.0 だと答えを選ぶことができないため 0.1 とした．この仕組みは，特定のキーワードに基づいた推論が不安定な場合に，他の確信度の高いパスがそれを補正する「自己修正能力」として機能する．また，集計されたスコアは最終判定の説明可能な信頼度として提示することが可能であり，単なるラベル出力に留まらない透明性の高い判定結果を提供する.

4 評価実験

本章では，提案手法の有効性およびキーワード抽出手法が語義判定に与える影響を検証するための実験設定について述べる.

4.1 実験設定

対象データセット　文書評価には，**FEWS(Few-shot Example for Word Sense Disambiguation)**[1] の Test split を用いた．FEWS は Wiktionary 由来の詳細な語義定義を基盤としており [10], 低頻度語を含む多様な多義語を網羅している点が特徴である. 本実験では, 事前の追加学習や提示事例 (Few-shot) を一切行わない Zero-shot 環境での性能を測定した.

使用モデルおよび環境　提案手法の各ステップにおける LLM の実装には, OpenAI 社の API を使用した.

- **Step2** : GPT-4.1-mini を使用．多様な表現を生成させるため，Tempereture は 0.2 に設定した.
- **Step3** : GPT-4-Turbo を使用．推論の安定性と再現性を確保するため，Temperature は 0.0 に設定した.

文脈キーワード抽出の設定　BM25 によるキーワード抽出には，FEWS で提供されている Sense Inventory を用いて構築したインデックスを使用した．これは，対象文中の単語が各語義を特徴づける記述とどの程度統計的に結びついているかを測定するためである．また，BM25 のパラメータの値は k_1=1.5，b = 0.75 とした．対象文からストップワードを除外した後，BM25 スコアの高い上位 5 単語を抽出し，推論のアンカーとした.

評価尺度　語義の出現頻度の偏りを考慮し，評価尺度には F1 スコアを用いた.

4.2 比較手法

本研究の目的である「キーワード抽出による文脈固定の効果」を相対的に評価するため，以下の手法と比較を行った.

1. **Baseline A**: GPT-4-Turbo に対し，思考過程を挟まずに対象文と語義定義のみを与え，直接語義 ID を選択させる手法．LLM の素の能力を測定する.
2. **Baseline B**: GPT-4-Turbo に対し，CoT の形で指示を与え，モデル内部の Attention のみで推論を行わせる手法.
3. **提案手法 (LLM，TF-IDF)**: キーワード抽出に BM25 ではなく，それぞれ，LLM 自身に「重要語を 5 つ選べ」と指示する手法，TF-IDF を使ってキーワード抽出をする手法．統計的手法と意

表 1 FEWS による F1 スコア (%) の結果

モデル	F1
SemEq-Large-Expert	72.2
ESR_{large}	75.8
Baseline A	45.8
Baseline B	56.0
提案手法 (LLM)	75.2
提案手法 (TF-IDF)	75.0
提案手法 (BM25)	**76.9**

味的手法の差異，統計的手法の精度ごとの差異を検証する．

既存の WSD モデル　以下の 2 手法は，FEWS データセットにおいて高い性能が報告されている代表的なニューラルモデルである．

4. **SemEq**[11]: 語義の分散表現と等値性判定を用いた手法．文脈と語義定義 (Gloss) の整合性を高次元ベクトル空間で計算する．
5. **ESR**[3]: 文脈と語義表現を高度に統合したモデル．FEWS 等のベンチマークにおいて，教師あり/Few-shot 設定で当時の SOTA を記録した．

5　結果と考察

5.1　定量的評価

表 1 に FEWS データセットを用いた各手法の F1 値を示す．提案手法は，既存の強力なニューラルモデルである ESR を上回る 76.9%という精度を達成した．特筆すべきは，SemEq や ESR が fine-tuning を前提とするのに対し，提案手法は Zero-shot 環境でこれを達成した点である．これは，LLM の潜在知識を「外部から適切に構造化された思考プロセス」で誘導することが効果的であることを実証している．しかし，語義決定におけるトークン数は 1 タスク当たり 2036 であった．これは，複数語義に対する独立した推論とその統合という段階的構造に起因するものであるが，将来的には語義候補の事前絞り込み等により削減可能である．

5.2　キーワード抽出手法の比較

本研究の重要な知見は，キーワード抽出に LLM 自身を用いるよりも，統計的な BM25 を用いる方が高精度であった点にある (+1.7 ポイント)．この要因として，LLM 特有の確証バイアスが挙げられる．LLM 自身にキーワードを選ばせた場合，モデルが既知の (あるいは頻用される) 語義に合致するように都合の良い単語を優先して選んでしまう傾向が確認された．一方，BM25 は文脈中の単語頻度と希少性に基づき，統計的に客観的な手がかりを「アンカー」として強制的に提示する．この統計的な客観性が LLM の推論を拘束することで，ハルシネーションを抑制し，より頑健な意味理解を実現したと考えられる．

5.3　透明性と精度のトレードオフの解消

従来の WSD 研究では，精度の追求 (ニューラルモデル) と解釈可能性 (ルールベース等) はトレードオフの関係にあった．しかし提案手法は，BM25 による証拠の抽出と LLM による類推の明示化という人間が検証可能なステップを踏むことで，既存 SOTA に並ぶ精度を維持しつつ，なぜその判断に至ったかを事後検証することができる透明性を獲得した．これは，信頼性が重視される実タスクにおける WSD の新たなパラダイムを提示するものである．

6　おわりに

本研究では，大規模言語モデル (LLM) の内部推論に依拠する従来のブラックボックスなアプローチに対し，人間の認知プロセスを模倣して思考を外部に構造化する新しい WSD パイプラインを提案した．

実験の結果，統計的指標 (BM25) による文脈のアンカリングと類推例文の生成を組み合わせることで，追加学習を一切行わない Zero-shot 環境でありながら，高度にチューニングされた既存のニューラルモデル (ESR 等) を上回る 76.9%の精度を達成した．特に，キーワード抽出において LLM 自身の選択よりも統計的客観性を優先することが，確証バイアスの抑制と推論の頑健性に向上に直結することを明らかにした．

本研究の最大の意義は，これまでトレードオフの関係にあると考えられていた「高い判別精度」と「プロセスレベルの解釈可能性」が，外部構造化ステップを介在させることで両立可能であることを実証した点にある．今後は，加重多数決における確信度評価のさらなる精緻化に加え，推論パス間での対立を解消する論理的プロセスの明示化に取り組み，人間がより深く信頼・検証できる**説明可能な意味理解システム**への発展を目指す．

参考文献

[1] Blevins T., Joshi M., and Zettlemoyer L. Fews: Large-scale, low-shot word sense disambiguation with the dictionary. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics.**, 2021. https://aclanthology.org/2021.eacl-main.36/.

[2] Robertson S.E. and Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In **Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 232–241. ACM, 1994.

[3] Yang S., Xin O., Hwee N., and Qian L. Improved word sense disambiguation with enhanced sense representations. **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 4311–4320, 2021.

[4] Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q.V., Zhou D., et al. Chain-of-thought prompting elicits reasoning in large language models. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.

[5] George A. Miller. Wordnet: A lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[6] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In **Proceedings of the 5th annual international conference on Systems documentation**, pp. 24–26. ACM, 1986.

[7] Blevins T. and Zettlemoyer L. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1006–1017, 2020.

[8] Deshan S., Nicholas M., and Julian H. Glossgpt: Gpt for word sense disambiguation using few-shot chain-of-thought prompting. In **Procedia Computer Science**, 2025.

[9] Jones K.S., Walker S., and Robertson S.E. A probabilistic model of information retrieval: Development and comparative experiments. **Information Processing & Management**, Vol. 36, No. 6, pp. 779–840, 2000.

[10] Meyer C.M. and Gurevych I. Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In **Electronic Lexicography**, chapter 13, pp. 259–291. Oxford: Oxford University Press, 2012.

[11] Wenlin Y., Xiaoman P., Lifeng J., Jinanshu C., Dian Y., and Dong Y. Bridging semantics between words and definitions via aligning word sense inventories, 2021. https://aclanthology.org/2021.emnlp-main.610.pdf.