LLMによって生成した語義定義文を用いた単語語義のクラスタリング

吉川　昭汰　　佐々木　稔
茨城大学大学院　理工学研究科　　茨城大学大学院　理工学研究科

概要

本研究では，大規模言語モデル（LLM）を用いて語義の定義文を生成し，それに基づいてクラスタ割当を再分類する語義誘導（WSI）の枠組みを提案する．本手法は，分布的類似性に基づく従来のクラスタリングに意味的情報を統合することで，クラスタ構造の保持とインスタンスレベルの対応の両立を図る．

SemEval-2010 および SemEval-2013 における実験の結果，提案手法は従来のクラスタリング手法および既存の WSI 手法を，構造指標（NMI，V-M）およびインスタンス対応指標（F-B^3，Fuzzy-F-B^3）の両面で上回った．特に，支配的語義バイアスを緩和し，少数語義の回復を改善できることが確認された．

また，モデル間比較および入力情報のアブレーションにより，分布的情報と判別的情報の両方が語義誘導に重要であることが示された．以上より，LLM による明示的な意味表現を統合することは，WSI における重要な課題に対する有効なアプローチであると結論づけられる．

1 はじめに

語義誘導（Word Sense Induction: WSI）は，ラベル付きデータを用いずに語の異なる意味を自動的に発見することを目的とした課題であり，語彙意味論における基盤的問題である．WSI は情報検索，機械翻訳，意味検索など多くの応用タスクを支える重要な技術であるが，語義分布の偏りや少数語義の回復困難といった問題により，依然として困難な課題である．

従来の WSI 手法の多くは，分布的類似性に基づく文脈表現とクラスタリングに依存している．しかし，これらの手法は支配的語義バイアスの影響を受けやすく，少数語義が多数語義のクラスタに吸収されてしまう傾向がある．また，分布的表現は意味の明示的な構造を持たないため，誘導されたクラスタの解釈や整合性にも限界がある．

近年，大規模言語モデル（LLM）は高度な意味理解と説明能力を示しており，自然言語による語義の定義文を生成することが可能である．しかし，この能力を WSI に体系的に組み込んだ研究はまだ十分ではない．

本研究では，LLM によって生成された語義定義文を用いてクラスタ割当を再分類する枠組みを提案する．本手法は，教師なしクラスタリングによって得られた初期クラスタに対して意味的な精緻化を施し，語義境界の明確化と少数語義の回復を同時に実現することを目指す．

本論文の貢献は以下の通りである．

(1) LLM による定義文を用いた再分類という新しい WSI 枠組みを提案する．

(2) 複数の評価指標およびデータセットにおいて本手法の有効性を実証する．

(3) 少数語義の回復および入力情報の役割に関する分析を行う．

2 関連研究

2.1 WSD

WSD は単語の正しい意味を決定するタスクで，辞書ベースや統計的手法，教師あり学習へと発展してきた．近年は Few-shot 学習や Zipf の法則を活用した手法で，低頻度語義への対応が改善されている．

2.2 WSI

WSI は教師なしで単語の意味をクラスタリングするタスク．言語モデルを用いた代替表現や相互情報最大化手法などにより，高精度な分類が可能になっている．論文では，WSI によるクラスタリングを行い，その結果を基に LLM で定義文を生成し，語義ごとの分類を行う．

2.3 LLM

Transformer（Vaswani ら）の登場以降，BERT や GPT など大規模言語モデルが急速に進化．GPT-4 や Google Gemini などはマルチモーダル対応や高度な推論能力を備え，幅広い NLP タスクで大きな性能向上を実現している．

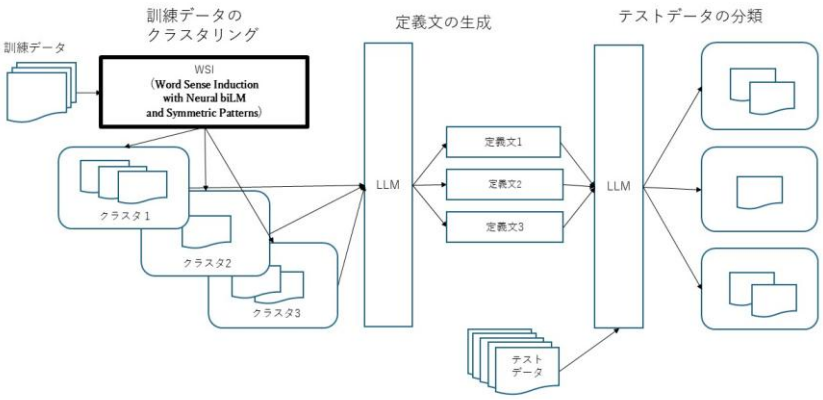

図 1　提案手法の概略図

2.4 Word Sense Induction with Neural biLM and Symmetric Patterns

ELMo による双方向 RNN から得る単語分布を，S-CODE 上に埋め込んでクラスタリングする．代替語セットの TF-IDF 化と凝集型クラスタリングにより柔軟に語義を推定し，ソフトクラスタリングで最終的な分類精度を向上させる．本研究は生成 AI による定義文生成を併用し，さらなる改善を図る．

本研究では，WSI-NS の枠組みに大規模言語モデルによる語義定義文生成を統合し，少数派語義を含む全語義をより公平に扱う WSI 手法を提案する．SemEval データセットでの評価を通じて，その有効性を検証する．

3 提案手法

本研究では，語義誘導において生じる多数派語義への偏り（majority sense bias）を緩和するため，WSI と大規模言語モデル（LLM）を統合した語義誘導手法を提案する．提案手法は，(1) 初期クラスタリング，(2) LLM による語義定義文の生成，(3) 定義文に基づくテストデータ分類の 3 段階から構成される．

図 1 に概略図を示す．学習データはまず，WSI によってクラスタリングされ，次に各クラスタに対して，LLM が対応する語義定義を生成する．これらの生成された定義は，その後，未知のテストデータを適切な語義クラスタに分類するために用いられる．

3.1 初期クラスタリング

SemEval-2010 および SemEval-2013 の XML 形式データセットを用い，対象語の用例文を教師なしでクラスタリングする．各 instance は文脈テキストと品詞情報を含むが，語義ラベルは付与されていない．

初期クラスタリングには WSI-NS を用いる．まず，ELMo の双方向言語モデル（biLM）により各用例文を符号化し，文脈依存の埋め込み表現を得る．次に，″X and other Y″ などの対称パターンから意味的に類似する代替語を抽出し，TF-IDF ベクトルとして表現する．これら 2 種類の特徴量を統合し，凝集型クラスタリングにより語義クラスタを推定する．さらに，語義境界の曖昧性に対応するため，各インスタンスから複数の代替語を生成し，それらのクラスタ割当の度数分布をインスタンスの語義所属の重みとして用いる．これにより，各用例は複数クラスタへの部分的な所属を許され，ソフトクラスタリングとして扱われる．

3.2 LLM による語義定義文の生成

初期クラスタリングで得られた各クラスタに対して，LLM を用いて語義定義文を生成する．これは，少数派語義の意味的特徴を明示化し，後段の分類における識別手掛かりとするためである．

LLM への入力には，(1) クラスタ識別用 SVM の重みから抽出した特徴語，(2) 対象語との共起に基づく PMI 特徴語，(3) クラスタ内の代表用例文を用いる．LLM はこれらの情報を基に，クラスタが共有する意味を抽象化し，1 文の英語定義として生成する．

プロンプトは，Sense ID，SVM 特徴語，PMI 特徴語，代表用例文を入力とし，「クラスタの主要な意味を表す簡潔な 1 文定義を生成せよ」という指示から構成される．

3.3 テストデータの分類

生成した語義定義文を用いて，未知のテスト用例

に対する語義分類を行う．各テストインスタンスは対象語とその文脈から構成され，語義ラベルは与えられない．

分類時には，すべての語義定義文とテスト文を LLM に入力し，「最も意味的に適合する語義番号を選択せよ」という形式で語義を選択させる．これにより，クラスタ中心との距離のみに基づく従来手法とは異なり，意味内容に基づく柔軟な判定が可能となる．特に，出現頻度の低い語義であっても，定義が明確であれば適切に識別される可能性が高まる．

LLM の出力は語義番号として取得され，SemEval の gold key と比較することで分類性能を評価する．

4 実験

4.1 データセット

評価には SemEval-2010 Task 14 および SemEval-2013 Task 13 を用いた．両データセットはいずれも XML 形式で提供され，対象語と文脈のみが与えられる教師なし設定である．SemEval-2010 は比較的粗い語義区別を，SemEval-2013 はより細粒度でロングテール性の強い語義分布を含み，majority sense bias に対する頑健性の評価に適している．

4.2 比較手法

比較対象は k-means，HDP，WSI-NS，one-cluster baseline（1cpl）とした．特に 1cpl は近年強力な単純ベースラインとして知られており，本研究ではこれを下限ではなく重要な比較対象として位置づける．

4.3 実験設定

用例集合を学習用と分類用に分割し，比率を変えた複数条件で評価を行った．k-means および HDP の最大クラスタ数は 4，WSI-NS および提案手法は 8 とした．分類段階には GPT-5，GPT-4o，Gemini-2.5-Flash を用い，各条件を複数回実行して平均性能を報告した．評価指標は F-NMI，Fuzzy F-score of B^3，V-measure，Paired F-score，NMI を用いた．Paired F-score（PF-S）は，どの用例同士が同じ語義としてまとめられたかという関係の正確さを，インスタンス対単位で評価する指標である．

5 結果

5.1 LLM の比較

提案手法で用いる LLM に GPT-５を用いた場合を表１，GPT-4o を用いた場合を表２，Gemini-2.5-flash を用いた場合の表３にそれぞれのスコアを示す．

なお，表中の『Model』列の数値（x/y）は，データ全体のうち定義文生成に使用したインスタンスの割合（x 割）と，分類評価に用いた割合（y 割）をそれぞれ示している．

SemEval-2013 では Gemini-2.5-Flash が最も高い Fuzzy-F-B^3 を示し，曖昧で重なり合う語義の扱いに優れていた．一方，SemEval-2010 では GPT-5 が最も高い F-B^3 を達成し，明確な語義境界の識別に適していることが示された．GPT-4o は両タスクで相対的に低い性能であった．

表 1　GPT-5 を用いたときの結果

Model	SemEval 2013		SemEval 2010			
	F-NMI	F-F-B^3	V-M	PF-S	NMI	F-B^3
1 / 9	20.09	60.27	41.52	60.85	41.52	66.06
2 / 8	14.11	54.50	44.33	62.23	44.33	67.17
3 / 7	15.08	55.17	44.56	63.65	44.56	68.13
4 / 6	14.65	55.13	46.45	64.52	46.45	69.34
5 / 5	14.60	56.57	47.52	64.22	47.52	70.26
6 / 4	14.47	56.37	49.57	64.59	49.57	71.85
7 / 3	17.22	58.71	52.00	64.14	52.00	73.30
8 / 2	17.83	58.73	55.96	62.76	55.96	76.39
9 / 1	16.59	63.07	59.34	54.71	59.34	80.02
10 / 10	13.86	54.78	46.26	64.67	46.26	68.68

表 2　GPT-4o を用いたときの結果

Model	SemEval 2013		SemEval 2010			
	F-NMI	F-F-B^3	V-M	PF-S	NMI	F-B^3
1 / 9	15.05	55.72	15.57	39.27	15.57	46.89
2 / 8	12.16	50.81	16.01	42.76	16.01	48.95
3 / 7	12.91	51.15	17.39	45.10	17.39	51.19
4 / 6	12.05	50.87	20.10	45.37	20.10	52.22
5 / 5	12.09	51.60	24.29	46.44	24.29	54.54
6 / 4	11.46	51.50	28.67	48.18	28.67	57.84
7 / 3	13.49	53.11	33.53	47.96	33.53	60.90
8 / 2	13.83	53.70	41.96	49.02	41.96	67.43
9 / 1	13.08	57.71	50.54	41.76	50.54	73.45
10 / 10	21.57	61.03	13.81	47.43	13.81	52.87

表 3　Gemini-2.5-flash を用いたときの結果

Model	SemEval 2013		SemEval 2010			
	F-NMI	F-F-B³	V-M	PF-S	NMI	F-B³
1 / 9	19.92	59.81	33.30	55.71	33.30	60.82
2 / 8	15.21	55.00	36.98	59.70	36.98	63.90
3 / 7	16.38	55.87	37.77	62.13	37.77	65.97
4 / 6	15.46	56.61	40.56	62.15	40.56	66.46
5 / 5	14.65	56.61	42.76	62.92	42.76	68.30
6 / 4	16.64	57.82	44.21	62.40	44.21	69.33
7 / 3	16.71	57.99	47.87	63.03	47.87	71.36
8 / 2	17.97	58.52	53.22	62.19	53.22	75.92
9 / 1	17.81	63.68	54.51	52.37	54.51	78.17
10 / 10	14.25	55.28	35.75	60.36	35.75	63.90

5.2 ベースラインとの比較

表 4 に提案手法は，k-means，HDP，WSI-NS，および 1cpl をいずれのデータセットにおいても上回るか同等以上の性能を示す．なお表中の提案手法の分割数はすべて 9/1 を示している．

1cpl は高い F-B³ を示すが語義区別を行わないため構造指標が 0 となる．一方，WSI-NS は構造指標が高いがインスタンス対応が弱いという傾向が確認された．

表 4 ベースラインとの比較

Model	SemEval 2013		SemEval 2010			
	F-NMI	F-FB³	V-M	PF-S	NMI	F-B³
k-means	16.94	31.05	24.86	47.04	24.86	52.92
HDP	8.98	58.13	11.8	54.51	11.89	58.13
1cpl	0	58.07	0	60.68	0	61.96
GPT-5	16.59	63.07	59.34	54.71	59.34	80.02
GPT-4o	13.08	57.71	50.54	41.76	50.54	73.45
Gemini-2.5-flash	17.81	63.68	54.51	52.37	54.51	78.17
WSI-NS	17.93	54.81	82.56	64.35	82.56	55.21

6 考察

6.1 定義文駆動型精緻化の効果

LLM による定義文を用いた再分類は，分布的類似性のみでは捉えきれない意味情報を補完し，表 4 の F-FB³，F-B³のスコアが向上し，正解語義との対応を大きく改善したことがわかる．また，少数語義の回復において有効であることが確認された．例えば，SemEval-2010 の対象語 chip.n において，電子部品としての「半導体チップ」を表す少数語義（sense 13）は，全体の出現頻度が極めて低く，従来のクラスタリング手法では木片や食品の chip と同一クラスタに吸収され誤分類される傾向があった．実際，従来手法では instance chip.n.111 は誤分類されていたのに対し，本研究の手法では「半導体材料としてのチップ」という定義文をアンカーとして用いることで正しく分類することができた．この結果は，定義文による意味的制約が，分布的に不利な少数語義を安定したクラスタとして保持する上で有効であることを示している．

LLM による定義文が有効であった理由は，文脈埋め込みが主に分布的類似性に基づく使用上の意味を捉えるのに対し，定義文は語義の内包的意味を明示的に言語化する点にある．文脈が似通った異なる語義は biLM では分離が困難であるが，定義文は上位概念や機能的役割といった概念的制約を付与することで語義境界を明確化し，少数語義が支配的語義に吸収されることを抑制すると考えられる．

6.2 モデル差の解釈

表１，３，４より，Gemini-2.5-Flash は連続的・曖昧な意味類似性の扱いに強く，GPT-5 はより明確な意味境界の識別に適している可能性がある．また，GPT-5 が最も総合的に高い水準のスコアを出すことができた．ただし，この差異の要因についてはさらなる検証が必要である．

6.3 少数語義のクラスタリング

提案手法は表４より NMI と F-B³の両方を高水準で改善し，少数語義を独立したクラスタとして保持しつつ，対応インスタンスを適切に回収できることが示された．

7 まとめ

本研究は，LLM による語義定義文生成と再分類を組み合わせた語義誘導手法を提案した．評価の結果，提案手法は構造保持指標とインスタンス対応指標の両方においてバランスよく高い性能を示し，特に少数語義の回復において有効であることが確認された．分布的類似性と判別的特徴を統合した定義文生成は，従来の WSI の限界を補完する有望な方向性を示す．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report.
arXiv preprint arXiv:2303.08774, 2024.

[3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Sorber, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models.
arXiv preprint arXiv:2312.11805, 2023.

[4] Asaf Amrami and Yoav Goldberg.
Word sense induction with neural biLM and symmetric patterns.
Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

[5] Asaf Amrami and Yoav Goldberg.
Towards better substitution-based word sense induction.
arXiv preprint arXiv:1905.12598, 2019.

[6] David Jurgens and Ioannis Klapaftis.
SemEval-2013 task 13: Word sense induction for graded and non-graded senses.
Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013), 2013.

[7] Adam Kilgarriff.
How dominant is the commonest sense of a word?
Proceedings of the 5th International Conference on Text, Speech and Dialogue, 2004.

[8] George Kingsley Zipf.
Human Behavior and the Principle of Least Effort.
Addison-Wesley Press, 1949.

[9] Hinrich Schütze.
Automatic word sense discrimination.
Computational Linguistics, Vol. 24, No. 1, pp. 97–123, 1998.

[10] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei.
Hierarchical Dirichlet processes.
Journal of the American Statistical Association, Vol. 101, No. 476, pp. 1566–1581, 2006.

[11] Terra Blevins and Luke Zettlemoyer.
Moving down the long tail of word sense disambiguation with gloss informed bi-encoders.
Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.