比喩検出における大規模言語モデルを用いた前後補助文脈の活用

林拓哉 [1]　佐々木稔 [1]
[1] 茨城大学大学大学院
{24nm752s, minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

比喩検出は、文字通りには解釈できない比喩表現を検出するタスクであり、文脈情報が重要である。過去の研究では、ChatGPT を用いて生成した補助文をターゲット文の前に追加する手法が提案され、精度向上が示されたものの、十分ではなかった。本論文では、ターゲット文の前後に補助文を追加する手法を提案し、比喩検出の精度向上を図った。複数のデータセットを用いた実験では、補助文の追加により精度、適合率、F1 スコアが過去の研究より向上したことが確認された。本研究は補助文生成の有効性を示しており、今後はより多様な文脈構成や最新の生成モデルを活用する可能性を探る必要がある。

1　はじめに

1.1　研究背景

比喩検出とは、文章中の比喩表現を検出することであり、本研究では英文の隠喩を対象とする。比喩表現は文字通りには解釈できず、その意味を推測する必要がある。近年、事前学習済み言語モデルを活用した Transformer を用いる手法が精度向上に寄与している [1][2]。特に MisNet モデル [3] は、1 文に基づいて比喩の有無を判断するが、文脈が不足する文では依然として困難が残る。

1.2　補助文脈の活用

過去の我々の研究 [4] では、ChatGPT を用いて補助文を生成・追加することで文脈情報を増強し、比喩検出の精度を向上させる手法を提案した。実験では補助文を追加したデータセットが元のデータセットよりも精度が高いことが確認されたが、文脈量の増加や追加位置の影響についても課題が示唆された。

1.3　研究目的

本論文では、補助文の追加位置を前後に変更し、文脈量を増加させた手法の有効性を検証する。

2　補助文を用いた比喩検出手法

本研究では、ある文（ターゲット文）およびその中の特定の単語（ターゲット単語）が比喩表現として使用されているかを判定するために、補助的な文脈を生成してターゲット文に付加する手法について述べる。 これまでの研究 [4] では、前に追加する前提で文章を生成したが、その際に作成した前に生成した文を結合した文に対して、後ろの文に補助的な文脈を生成させ、結合させる。

2.1　補助文脈の生成

ChatGPT を用いてターゲット文に続く補助文を生成する。モデルは [4] に合わせて、ChatGPT-3.5 Turbo を使用する。この際、2 種類のプロンプトを使用する。

プロンプト 1(Prompt1)　以下に 1 種類目のプロンプトを示す。プロンプト中の太字の単語は変数を表しており、それぞれの変数とその内容を表 1 に示す。

- 文章作成のプロンプト (Prompt1):
  "Generate a **N**-words sentence that **Verb**s "**sentence**" and in which '**target_word**' in "**sentence**" is **m** used as a metaphor."

プロンプト 2(Prompt2)　2 種類目のプロンプトは以下の通りである。文生成の基本的な指示はプロンプト 1 と同じであるが、プロンプト 1 の前に、ターゲット単語が比喩表現として使用されている例文と使用されていない例文を追加している。それぞれの変数は表 1 に示したプロンプト 1 と同じである。

表 1　プロンプトの変数

変数名	内容
N	追加される文の単語数
Verb	前に追加するときは “precede”、 後ろに追加するときは “follow”
sentence	ターゲット文
target_word	ターゲット単語
m	比喩表現が含まれる場合は空白文字、 含まれない場合は「not」

- 例文を含む文を生成するプロンプト（Prompt2）：
  "’derive’ in "For the moment let us use the above expression for deriving Biot-Savart’s law." is used as a metaphor."
  "’derive’ in "The next section therefore attempts to summarize what we do know; it is derived chiefly from the Earth System Sciences Committee (ESSC) (1988)." is not used as a metaphor."
  "Generate a **N**-words sentence that **Verb**s "**sentence**" and in which ’**target_word**’ in "**sentence**" is **m** used as a metaphor."

2.2　文の結合

前節で生成した文とターゲット文を結合する方法について説明する。この結合は「ターゲット文 + 生成された文」の形式で単純に配置する。

3　DataSet

3.1　元データセット

本節では、第 2 節で述べた通り、ターゲットの文は [4] で作成したデータセットを使用する。その実験で使用したデータセット及び構造について説明する。実験では、MOH-X[5]、VUA_All[6]、および VUA_Verb[6] の 3 種類のデータセットを使用した。

MOH-X　MOH-X は、動詞に焦点を当てたデータセットであり、比喩的意味と文字通りの意味の両方を含む動詞の用例が WordNet から収集されている。

VUA_All　VUA とは、アムステルダム自由大学（Vrije Universiteit Amsterdam）および同大学で作成された VU アムステルダム比喩コーパス（VU Amsterdam Metaphor Corpus）を指す。VUA は、BNC Baby Corpus から収集されたテキストを基にしており、学術、会話、小説、ニュースの 4 つのジャンルを含む。ターゲット単語にはすべての品詞が含まれている。

VUA_Verb　VUA_Verb は、VUA_All のうちターゲット単語が「VERB」とラベル付けされたデータである。

表 2　各データセットの情報 #Sent. は文数、#Target はターゲット単語数、%Met. は比喩を含む割合、Avg. Len は文の平均長を示す。

DataSet	#Sent.	#Target	%Met.	Avg. Len
VUA_All_{tr}	6,323	116,622	11.19	18.4
VUA_All_{val}	1,550	38,628	11.62	24.9
VUA_All_{te}	2,694	50,175	12.44	18.6
VUA_Verb_{tr}	7,479	15,516	27.9	20.2
VUA_Verb_{val}	1,541	1,724	26.91	25.0
VUA_Verb_{te}	2,694	5,873	29.98	18.6
$MOH-X$	647	647	48.69	8.0

3.2　MisNet のデータ形式

第 4.2 節で述べる通り、今回の実験は MisNet[3] を用いる。そのため、MisNet のデータ形式について説明する。文献 [3] では、データは csv 形式で提供されており、これらのデータセットを MisNet で使用するために再フォーマットされている（表 3 参照）。

表 3　MisNet のデータ形式

列名	内容
sentence	ターゲット文
label	比喩の有無を示すラベル 0: 含まれない 1: 含まれる
target_position	ターゲット単語の文中での位置
target_word	ターゲット単語
pos_tag	ターゲット単語の品詞
gloss	ターゲット単語の定義
eg_sent	ターゲット単語の例文

4　実験

4.1　追加の補助的文脈を含むデータセット

各文に対して、第 2 節で説明されている方法を使用して文脈補足文を生成し、元の文と生成された文を組み合わせて文脈補足文を作成する。本研究では、生成された文を MisNet データセットの形式に統合している。

4.1.1 生成手順

追加の補助文脈を含むデータセットを作成する手順について述べる。まず、MOH-X、VUA_All、またはVUA_Verbのデータセットを読み込み、これに基づいてプロンプトを作成する。次に、ChatGPTを使用して補助文を生成し、生成された文をターゲット文と結合する。最後に、データをMisNetデータセット形式にまとめ、csv形式で出力する。本研究では、ChatGPTが使用するトークン数をプロンプト1では20、プロンプト2では50に設定している。これらは第2節で述べたプロンプト内の変数と対応しており、プロンプトのsentenceおよびtarget_wordには、データセットのsentenceおよびtarget_wordが対応する。

4.2 実験方法

Transformerを用いた比喩検出モデルMisNet[3][1)]を、各データセットにおいて以下の3つの状況で実行した。(1) 何も追加しない場合、(2) 第2節で示したプロンプト1によって生成された補助文脈を追加した場合、(3) プロンプト2によって生成された補助文脈を追加した場合、という3つの状況である。それぞれのデータセット内のすべての文について、比喩の有無を予測し、予測ラベルを出力として取得した。本実験では、事前学習済みのBERTモデルとしてRoBERTaの基本モデル[2)]を使用した。

5 結果と考察

5.1 実験結果

5.1.1 元のラベルと予測ラベルの精度

以下に、元のデータセットおよび補足文を追加したデータセットにおける元のラベルと予測ラベルの精度を示す。太字の数字は各行の各項目における最大値を表している。行名P1およびP2は、それぞれプロンプト1およびプロンプト2を使用した文を示し、5wordsおよび10wordsは、前に5単語または10単語の文を追加したことを示している。数値内の太字は、各列における最大値を示す。全体的に、補助文が追加された場合に各スコアが元のデータよりも高い傾向がある。特に、正解率とF1スコアは全て元データより今回の手法を使用した方が高かった。

1) MisNetの詳細は付録にて説明する。
2) RoBERTaのウェブサイト https://huggingface.co/FacebookAI/roberta-base

表4 MOH-Xにおける
正答率、適合率、再現率、およびf1スコアの平均

	Acc	Prec	Rec	F1
original data	0.828565	0.819046	0.844503	0.827488
P1_5words	0.851078	0.854915	0.839839	0.844435
P1_10words	0.84529	0.820463	**0.88696**	**0.84934**
P2_5words	**0.85295**	**0.8601**	0.835703	0.846013
P2_10words	0.838578	0.824779	0.858565	0.838502

表5 VUA_Allのテストデータにおける
正答率、適合率、再現率、およびf1スコアの平均

	Acc	Prec	Rec	F1
original data	0.940933	0.771933	**0.75507**	0.761333
P1_5words	**0.94886**	0.826496	0.745475	**0.7839**
P1_10words	0.947683	0.827125	0.732661	0.777032
P2_5words	0.947922	0.821923	0.742271	0.780069
P2_10words	0.94854	**0.83036**	0.736985	0.780889

表6 VUA_Allの検証データにおける
正答率、適合率、再現率、およびf1スコアの平均

	Acc	Prec	Rec	F1
original data	0.940933	0.771933	0.755067	0.761333
P1_5words	**0.95172**	0.806882	**0.76822**	**0.78708**
P1_10words	0.951149	**0.8122**	0.753733	0.781875
P2_5words	0.950735	0.799768	0.768219	0.783676
P2_10words	0.951201	0.811542	0.755293	0.782408

表7 VUA_Verbのテストデータにおける
正答率、適合率、再現率、およびf1スコアの平均

	Acc	Prec	Rec	F1
original data	0.802133	0.679867	**0.7374**	0.6972
P1_5words	0.840116	0.748489	0.70301	0.725037
P1_10words	**0.84454**	**0.76566**	0.693924	0.728031
P2_5words	0.841818	0.740462	0.727428	**0.73389**
P2_10words	0.836881	0.739416	0.704145	0.72135

表8 VUA_Verbの検証データにおける
正答率、適合率、再現率、およびf1スコアの平均

	Acc	Prec	Rec	F1
original data	0.8112	0.653267	0.760133	0.6932
P1_5words	0.866589	0.75	0.756466	0.753219
P1_10words	**0.87587**	**0.78409**	0.743534	**0.76327**
P2_5words	0.867169	0.748414	**0.76293**	0.755603
P2_10words	0.862529	0.750552	0.732759	0.741549

5.1.2 t 検定による優位性の評価

各データセットについて、予測ラベルの平均値が 0.5 未満の場合は 0、0.5 を超える場合は 1 とするデータを作成した。帰無仮説は、「状況の異なる 2 つの場合について、結果に有意な差がないこと」であり、t 検定を実施した。P(T<=t) 両側値は以下の通りである。テストデータの結果について述べる。

MOH-X では、値が 0.1%未満であった 2 つのケースは、元のデータとプロンプト 1 の前に追加の 10 単語の文を加えた場合のみである。他のケースでは大部分が 0.2 から 0.9 の範囲であり、有意水準の 5%または 1%を下回るものはなかった。

VUA_All では、プロンプト 1 で前後に 5 単語ずつ追加したデータとプロンプト 2 で 5 単語ずつ追加したデータの場合、プロンプト 2 で前後に 5 単語ずつ追加したデータとプロンプト 2 で前後に 10 単語ずつ追加したデータの場合が有意水準を上回った。しかし、それ以外は全て 10%の優位水準を下回り、1%の優位水準を下回るケースが多い。

VUA_Verb では、過半数が 1%の優位水準を下回っている。

MOH-X	P1_5words	P1_10words	P2_5words	P2_10words
元データ	0.27664009	0.056698671	0.24851806	0.41463747
P1_5words		0.330774043	0.87332708	0.723968172
P1_10words			0.28919413	0.206148785
P2_5words				0.752098726
P2_10words				

図 1　MOH-X の t 検定結果 ($P(T \leq t)$ 両側)

VUA_All(test)	P1_5words	P1_10words	P2_5words	P2_10words
元データ	3.1813E-31	1.27587E-22	3.3693E-36	1.48251E-43
P1_5words		0.050357514	0.19533223	0.009797787
P1_10words			0.00095415	5.50912E-06
P2_5words				0.182792346
P2_10words				

図 2　VUA_All のテストデータ t 検定結果 ($P(T \leq t)$ 両側)

VUA_All(vald)	P1_5words	P1_10words	P2_5words	P2_10words
元データ	1.32165E-16	8.35281E-13	2.20961E-11	1.715E-19
P1_5words		0.265044224	0.115539794	0.403163708
P1_10words			0.679528876	0.050916568
P2_5words				0.020856356
P2_10words				

図 3　VUA_All の検証データ t 検定結果 ($P(T \leq t)$ 両側)

5.2　考察

第 5.1.1 節の結果から、補助文を追加することで比喩検出の精度が向上することが確認された。特に適合率の向上により、比喩がある文を正しく予測する能力が向上した。しかし、再現率が低下しており、比喩がある文を比喩なしと誤予測する傾向が示唆された。これは、真陽性の増加に比べて偽陰性の増加が大きかったためと考えられる。また、t 検定の結果から、文脈を追加した効果を明確に判定するのは難しく、多くの場合有意水準を下回っていた。ただし、元データと比較すると 1%の有意水準を達成しており、文脈の追加が効果を持つことは示されている。一方で、プロンプト 1 とプロンプト 2 の間に明確な差は見られず、文字数の違いについての効果も今後の課題といえる。特に、動詞に注目したメタファー検出では効果が限定的であり、動詞の多義性や文脈に応じた使い分けの多さが原因と考えられる。また、動詞のすべての意味を網羅しているわけではないため、正確な意味認識が難しいことも一因である。これらの課題を踏まえ、動詞におけるメタファー検出の改善が今後の重要な研究課題である。

VUA_Verb(test)	P1_5words	P1_10words	P2_5words	P2_10words
元データ	0.39199084	0.000344642	3.3764E-07	0.612131445
P1_5words		0.005011174	8.9381E-06	0.16304327
P1_10words			0.16029551	3.64863E-05
P2_5words				3.40195E-08
P2_10words				

図 4　VUA_Verb のテストデータ t 検定結果 ($P(T \leq t)$ 両側)

VUA_Verb(vald)	P1_5words	P1_10words	P2_5words	P2_10words
元データ	0.412779976	0.358945802	0.117554108	0.048275724
P1_5words		0.075040538	0.015271038	0.436859367
P1_10words			0.491455928	0.332119881
P2_5words				0.098973712
P2_10words				

図 5　VUA_Verb の検証データ t 検定結果 ($P(T \leq t)$ 両側)

6　まとめ

本論文では、比喩検出の精度を向上させる手法において、補助的な文脈を前と後ろに追加した場合の有効性について検証した。提案手法を用いて補助文を追加したデータセットと元のデータセットについて、比喩検出モデルである MisNet を用いた予測結果を比較した。その結果、比喩検出の精度が向上することが示された。特に、補助的な文脈を前に追加した場合よりも、全体的に精度が向上した。今後の課題としては、文の後ろのみに追加した場合や補助的な文脈の単語を増やす場合が挙げられる。また、使用する ChatGPT モデルを最新モデルにした際の結果も調査したい。

参考文献

[1] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1763–1773, Online, 2021. Association for Computational Linguistics.

[2] Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. Verb metaphor detection via contextual relation learning. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4240–4251, Online, 2021. Association for Computational Linguistics.

[3] Shenglong Zhang and Ying Liu. Metaphor detection via linguistics enhanced siamese network. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 4149–4159, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.

[4] Takuya Hayashi and Minoru Sasaki. Metaphor detection with additional auxiliary context. In **2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 121–126, 2024.

[5] Saif Mohammad, Ekaterina Shutova, and Peter Turney. Metaphor as a medium for emotion: An empirical study. In **Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics**, pp. 23–33, Berlin, Germany, 2016. Association for Computational Linguistics.

[6] Gerard Steen. **A Method for Linguistic Metaphor Identification: From MIP to MIPVU**, Vol. 14. John Benjamins Publishing, 2010.

A　参考情報

本節には，本文にて説明しきれなかった技術について説明する。

A.1　MisNet

A.1.1　MisNet とは

MisNet (Metaphor Identification from Siamese Network) は、Shenglong Zhang と Ying Liu による論文 "Metaphor Detection via Linguistics Enhanced Siamese Network" [3] [3)] で提案された比喩検出モデルである。このモデルは、MIP と SPV の 2 つの言語規則に基づいて計算を行い、その結果を合算して比喩を含むかを判定する。MisNet はいくつかのデータセットで従来のモデルより優れた性能を示した。

MIP とは　MIP (Metaphor Identification Procedure) は、比喩を識別する手続きを示す。この研究では、文脈的な意味と対象単語の基本的な意味の類似度を計算する。基本的な意味は辞書から取得する。

SPV とは　SPV (Selectional Preference Violation) は、選択的優先違反を指し、単語が文脈で一般的に使用されない状況を表す。この研究では、対象単語と文脈の不一致度を計算する。

MisNet の構造　MisNet の構造は図 6 に示される。右側の入力は表 9 の形式で BERT エンコーダーに入力され、その出力は h_{MIP} と h_{POS} の計算に使用される。左側の入力は表 10 の形式で同じ重みを共有する BERT エンコーダーに入力され、h_{SPV} と h_{MIP} の計算に使用される。特徴タグ GF、LF、POS、TAR は、それぞれグローバル特徴、ローカル特徴、品詞特徴、ターゲット単語を示すものである。以下の式で h_{MIP}、h_{SPV}、h_{POS} を統合する。

$$y = \sigma(W^T[h_{MIP}; h_{SPV}; h_{POS}] + b) \tag{1}$$

ここで、W は重み、b はバイアス、σ はソフトマックス関数である。$y \in \mathbb{R}^2$ の値で比喩の有無を判別する。

表 9　右側の BERT への入力形式

先頭タグ	対象の文章	文末タグ
一行目の各単語の座標		
一行目の各単語の特徴タグ		

3)　https://github.com/silasthu/misnet にてプログラムが公開されている。

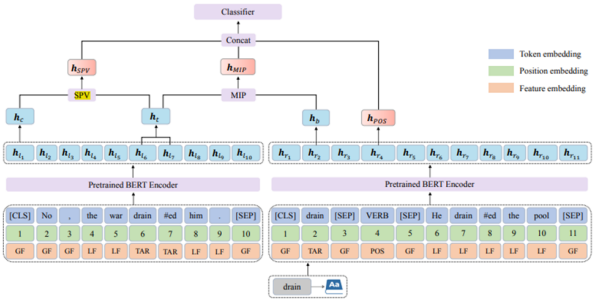

Figure 3: MisNet architecture. The two BERT encoders share weights. h_c, h_t, h_b are context embedding, contextual target meaning, and basic meaning respectively. GF, LF, POS, TAR denote global feature, local feature, POS feature and target word.

図 6　MisNet の構造 ([3] より引用)

表 10　左側の BERT への入力形式

先頭タグ	対象単語	文末タグ	品詞	文末タグ	対象の文	文末タグ
一行目の各単語の座標						
一行目の各単語の品詞タグ						

また、分類タスクでは、最適化基準としてクロスエントロピー損失を使用する。

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(\hat{y}_i) \tag{2}$$

上記の式の N はトレーニングサンプルの数である。y_i と $\hat{y}_i$ はそれぞれ i 番目のサンプルの正解ラベルと予測スコアを表す。w_{y_i} はクラスの重みで、データの不均衡問題を緩和するためのものである。

A.1.2　MisNet のトレーニング・Metaphor の判定

モデルのトレーニング　4.2 節で示した形式のトレーニングデータを抽出し、4.1 節で述べた構造のモデルに入力して (1) 式を計算する。その後、計算された変数 y とトレーニングデータの真のラベルとの差異を評価するために (2) 式で示す損失関数を適用する。この損失を用いて、勾配の計算が行われ、その勾配を元にモデルのパラメータを更新する。これらの手順を各エポックごとに繰り返す。

Metaphor の判定　未知の文から Metaphor の判定をする際は、トレーニングと同じようにデータを抽出し、Transformer を使用してロジットを計算する。その後、ロジットの各行において最大の値と対応したラベルを抽出する。このラベルが表 3 の label に相当する。