6W-06

小説テキストに出現する直喩表現の抽出手法の検討

宮脇　星名†　　安藤　一秋†

†香川大学

1. はじめに

比喩表現は，ある物事を別の事柄に例えることで，文字通りの表現以外の情報を表現・伝達する機能と詩的・審美的効果を喚起する機能の 2 つを持つ[1]．比喩は，主に「直喩」，「隠喩」，「換喩」，「提喩」の 4 つに分類される．直喩の判定においては，手がかりとして「のような」といった特定の表現（比喩指標）があるため，他の比喩より判定が容易であるといえる．田添らは「名詞 A のような名詞 B」表現について比喩性を判定する手法[2]を提案しており，新聞記事に対して 65.6%の性能で比喩性を判定できると述べている．事前検討として，田添らの手法を再現実装し，小説テキストに対して実験した結果，73.7%の判定性能が得られることを確認した．そして，誤判定の傾向を分析した結果，誤判定文には時間や人称代名詞など特定の語が含まれていることから，単純なルールベースで性能改善できる余地を確認した．

本研究では，小説テキストから比喩表現を自動抽出するシステムの構築を目的とする．本稿では，事前検討として，田添らが注目した「名詞 A のような名詞 B」に加え，「名詞 A のような名詞句 B」表現を対象に，単純なルールベースで直喩を抽出するモデルについて検討する．

2. 実験概要

本稿では，以下の 3 手法について検討する．

・手法 1

テキストから「のような」表現を検出し，前後が名詞かつ特定の名詞でないならば比喩と判定する．特定の名詞については，日本語 WordNet[3]から誤判定に関与している特定の語句群の類語と，デジタル大辞泉[4]と精選版日本国語大辞典[5]から代名詞を集め，589 語の判定辞書を構築して利用する．

・手法 2

手法 2 は，図 1 に示す田添らの再現手法に対し，手法 1 の判定辞書を用いて特定の名詞を除外する手法である．

・手法 3

田添らの手法は「名詞 A のような名詞句 B」表現を対象にしていない．そこで，このタイプの比喩を抽出するため，図 2 に示す抽出フローにより，形態素解析の品詞情報を用いてパターンを分類する．図 2 のステップ 2 の特定の品詞は，日本語係り受け解析器 Cabocha[6]で形態素解析した結果，名詞・形容詞・記号・連体詞・接頭詞に該当するものを指す．ステップ 3 の特定の品詞は，名詞・形容詞・記号（並列表現に使用するもののみ）・連体詞・接頭詞・接続詞（並列のみ）・助詞に該当するものを指す．本稿では，名詞句 B の終了条件として，ステップ 4 に示すように，動詞・助動詞・助詞（係助詞・格助詞）の出現と定めている．これにより，「祖母の口癖のような“他人に優しくする”事は……」といった記号で示される括弧内に名詞以外が出現するパターンを除外している．また，手法 1 と手法 2 で用いた判定辞書を使用し，名詞 A と名詞句 B 内に特定の名詞が出現するならば除外する．これらにより抽出された文に対して，田添らの再現手法を適用し，名詞 A と名詞句 B に含まれる名詞の意味情報を用いて，比喩を抽出する．

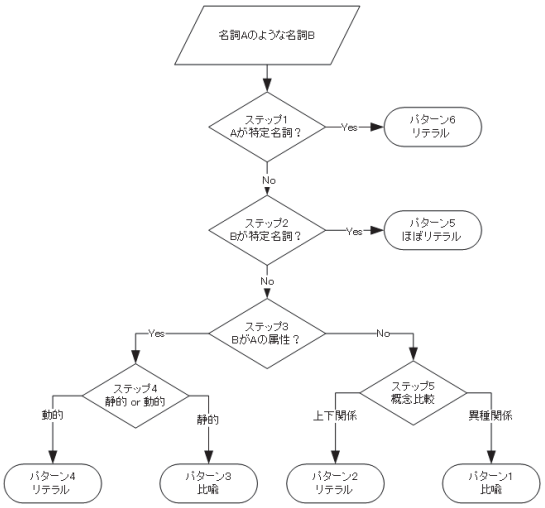

図 1　田添らの再現手法（文献[2]からの引用）

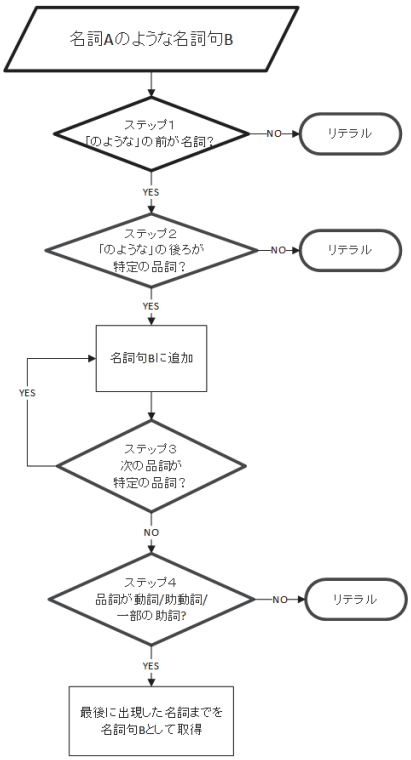

図 2　「名詞 A のような名詞句 B」の抽出フロー

An Extraction Method of Simile Expressions in Text of Novels

Seina Miyawaki †, Kazuaki Ando†

† Kagawa University

対象テキストには，「青空文庫」から 100 作品と「小説家になろう」の 6 ジャンル（ドラマ・歴史・ホラー・推理・ハイファンタジー・ローファンタジー）から任意抽出した各 100 作品を利用する．全文に対し，括弧に該当する記号を正規化したものを，各手法の入力文として利用する．表 1 に各ジャンルの総文数と総文字数を示す．

本実験では，小説記事 100 作品中の比喩抽出数を比較すると共に，手法 1 と手法 2，手法 3 が比喩として抽出した結果のうち，各 50 件を人手で正誤判定した精度により，各手法の性能を評価する．

表 1　使用した小説テキスト

小説ジャンル	総文数	総文字数
青空文庫	88,600	2,864,059
小説家になろう		
ドラマ	3,240,745	47,394,498
歴史	4,057,558	65,719,130
ホラー	1,380,975	20,489,859
推理	1,401,460	20,682,921
ハイファンタジー	12,837,842	194,727,498
ローファンタジー	6,678,372	93,221,979

3. 実験結果と考察

各作品から抽出した比喩表現を含む文数を表 2 に示す．表 1 から分かるように，青空文庫の総文数と総文字数が著しく少なく，ハイファンタジーの総文数と総文字数が一番多い．抽出された比喩表現もその総文数と総文字数に対して，比例傾向にある．

表 3 に，手法 1 と手法 2，手法 3 が比喩として抽出した結果のうち，各 50 文を人手で正誤判定した結果を示す．表 3 の(a)より，手法 1 の平均精度は 74.9%となり，田添らの再現手法の精度 73.7%と同程度となった．次に，表 3 の(b)より，手法 2 の平均精度は 92.6%であった．この結果から，田添らの再現手法に対して，単純なルールベースで精度が改善できたといえる．表 3 の(c)より，手法 3 の平均精度は 85.4%となり，「名詞 A のような名詞句 B」表現に対しても，田添らの再現手法と判定辞書が有効であったといえる．

手法 3 のエラー分析の結果，名詞句 B の区切りが正しくなく，名詞 A が名詞句 B に対して比喩として修飾していない誤判定が存在していた．例として，「冬のような冷たい朝の朝礼では……」が挙げられる．名詞句 B として「冷たい朝の朝礼」を検出しているが，「冬」は「朝」を修飾している．また，「柵のようなものと見張り台」では，名詞句 B として「ものと見張り台」を抽出しており，並列関係の誤りが見られた．

3 手法に共通している点として，「鈴のような女生徒」といった多義語による誤判定が存在していた．また，文脈に依存した比喩も同様に対処できていないことを確認した．これらは，ルールベースでの改善が難しいため，今後は，言語モデルを用いて比喩を抽出する手法について検討する．

4. おわりに

本稿では，「名詞 A のような名詞 B」表現に対して，田添らの再現手法をルールベースで改善する手法について検討した．その結果，田添らの再現手法と比較して，約 18.9 ポイントの精度向上を確認した．また，「名詞 A のような名詞句 B」表現を対象にする手法を検討し，85.4%の精度で比喩を抽出できることを確認した．

本稿では，ルールベースによる比喩抽出の性能改善を試みたが，単語の多義性の問題や文脈に依存する比喩などに対しては，ルールベースでの抽出に限界を確認した．

今後は，抽出性能の向上を目指して，言語モデルによる改善法を検討する．

表 2　各手法から抽出された比喩の文数

ジャンル	手法1	手法2	手法3
青空文庫	274	71	82
小説家になろう			
ドラマ	4,256	1,235	1,256
歴史	5,524	1,537	1,553
ホラー	2,547	767	801
推理	2,196	665	694
ハイファンタジー	22,856	6,332	6,512
ローファンタジー	9,918	2,776	2,943

表 3　各 50 文に対する正誤判定の結果

	青空文庫	ドラマ	歴史	ホラー	推理	ハイファンタジー	ローファンタジー
正	40	29	32	46	44	35	36
誤	10	21	18	4	6	15	14
精度	0.80	0.58	0.64	0.92	0.88	0.70	0.72

（a）手法 1 の結果

	青空文庫	ドラマ	歴史	ホラー	推理	ハイファンタジー	ローファンタジー
正	44	47	44	50	47	46	46
誤	6	3	6	0	3	4	4
精度	0.88	0.94	0.88	1.0	0.94	0.92	0.92

（b）手法 2 の結果

	青空文庫	ドラマ	歴史	ホラー	推理	ハイファンタジー	ローファンタジー
正	42	44	41	47	47	40	38
誤	8	6	9	3	3	10	12
精度	0.84	0.88	0.82	0.94	0.94	0.80	0.76

（c）手法 3 の結果

参考文献

[1] 内海 彰，“比喩によってどのような詩的効果が喚起されるか　比喩の鑑賞仮定の認知モデルに向けて”，第 17 回人工知能学会全国大会論文集，2003.
[2] 田添 丈博, 椎野 努, 桝井 文人, 河合 敦夫，“名詞 A のような名詞 B 表現の比喩性判定モデル”，自然言語処理，10 巻，2 号，pp.43-58，2003.
[3] 日本語 WordNet，http://compling.hss.ntu.edu.sg/wnja/
[4] デジタル大辞泉，https://kotobank.jp/dictionary/daijisen/
[5] 精選版日本国語大辞典，https://kotobank.jp/dictionary/nikkokuseisen/
[6]工藤 拓,松本 裕司，“チャンキングの段階適用による日本語係り受け解析”，情報処理学会論文誌，Vol.43，No.6，pp.1834-1842，2002.