E-025

小説テキストからの「ような」表現に基づく直喩表現抽出手法の検討
A Simile Extraction Method Based on “Youna” Expressions from Text of Novels

宮脇 星名† 安藤 一秋†
Seina Miyawaki Kazuaki Ando

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．その中でも直喩表現は，「ような」といった特定の比喩指標が使われる傾向にある．

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，小説テキストに出現する「ような」表現に着目し，「動詞 A ような名詞 B」に基づく直喩表現を抽出対象とする．

近年，機械学習を用いたアプローチが主流ではあるが，比喩の特徴把握を目的に，まずは規則のみを用いて直喩表現を抽出する手法について検討する．そして，規則に基づく直喩抽出の限界を確認すると共に，問題点を分析し，抽出性能を改善する手法について検討する．

2. 先行研究

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，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．我々の先行研究[3]において，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，この再現手法に対して，比喩ではない文（リテラル）に多く使われている特定の語句群を除外するルールを追加し，92.4%の判定性能が得られることを確認した．さらに，「名詞 A のような名詞句 B」表現に対して，新たな判定ルールを構築し，特定の語句を除外した結果 85.4%の判定性能が得られることを確認した．

3. 抽出実験

本稿では，「名詞 A のような名詞句 B」に続いて，「動詞 A ような名詞 B」タイプの直喩表現を抽出対象とする．

本実験では，先行研究[3]における「名詞 A のような名詞 B」タイプの比喩抽出手法を参考に設計した，2 種類の抽出手法を用いて比喩文の抽出性能を評価する．

・手法 1（ベースライン手法）

テキストから「ような」を検出し，図 1 に示す抽出フローに基づいて，形態素解析の品詞情報を用いてパターン分類し，比喩文を抽出する．

・手法 2

手法 1 に加えて，リテラルに多く使われている「する，いる，なる，ある，くる，みる，いう，もらう，くれる，できる，形容動詞語幹につく助動詞」といった非自立動詞などの特定の語句を含む文を除外することで比喩を抽出する．

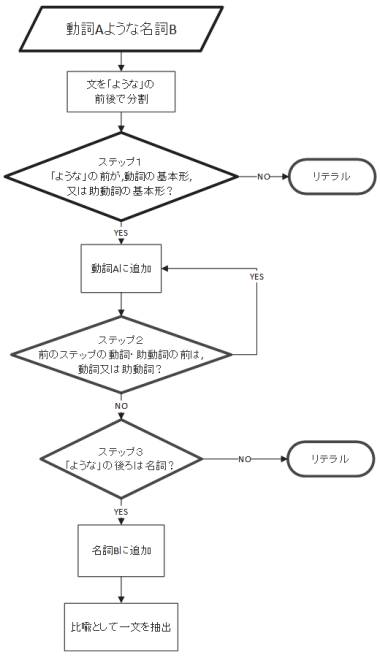

図 1　手法 1 に基づく抽出フロー

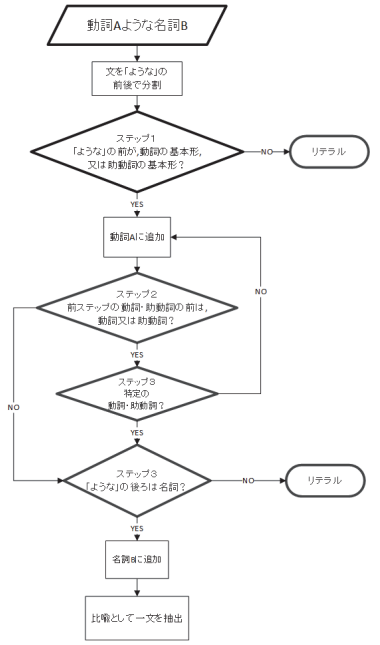

図 2　手法 2 に基づく抽出フロー

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．

† 香川大学 創造工学部，Faculty of Engineering and Design, Kagawa University

表 1 に各ジャンルの総文数と総文字数を示す．手法 1 と手法 2 では，小説記事 100 作品中の比喩抽出数を比較すると共に，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

4. 実験結果と考察

手法 1 と手法 2 で抽出した結果に対して，各 50 文を人手で正誤判定した結果を表 2 の(a)と(b)に示す．表 2 の(a)に示すように，手法 1 の平均精度は 12.29%であり，(b)に示すように，手法 2 の平均精度は 17.42%となった．手法 2 は，手法 1 より，精度が 5.13 ポイント向上したことから，特定の語句の除外については，ある程度の効果を確認した．しかし，どちらも 20%未満の精度になったことから「名詞 A のような名詞 B」では有効的だった単純な規則に基づく判定が困難であり，規則に基づく抽出手法の限界を確認した．

5. 追加実験

「動詞 A ような名詞 B」において，動詞 A の概念と名詞 B の概念の距離が遠いほど，比喩性が高いと考えらえる．そこで，機械学習に基づく抽出手法を検討するための事前検討として，動詞 A と名詞 B の単語分散表現の類似度を算出し，比喩文とリテラルにおける差の有無を確認する．

単語分散表現には，日本語 Wikipedia エンティティベクトル[4]，日本語 fastText モデル[5]，HR 領域のワードベクトル[6]，hottoSNS-w2v[7]，朝日新聞単語ベクトル[8]，chive[9]）用いる．

名詞と動詞の類似度については，手法 2 において人手で比喩だと判定した 100 文とリテラルだと判定した 100 文を比喩でない文として利用する．なお，未知語については，モデルによって数文が除外となっている．

実験結果を表 3 に示す．表 3 より，すべての単語分散表現において，比喩文の動詞 A と名詞 B の単語間の類似度が小さく，比喩でない文の類似度が大きいことを確認した．特に，⑤朝日新聞単語ベクトルの glove-re を用いた場合，差が最大となり，0.1010 となることを確認した．したがって，比喩文とリテラルにおいて，動詞と名詞の類似度の違いから，比喩を判定できる可能性が確認できた．

今後は，この特徴を利用した機械学習モデルで，「動詞 A ような名詞 B」タイプの直喩表現を抽出する手法について検討する．

6. おわりに

本稿では，「動詞 A ような名詞 B」タイプの直喩表現に対して，規則に基づく 2 つの抽出法について検討した．実験の結果，手法 1 では 12.29%，手法 2 では 17.42%の抽出精度となり，規則に基づく抽出手法の限界を確認した．そこで，「動詞 A ような名詞 B」において，動詞 A の概念と名詞 B の概念の距離が遠いほど，比喩性が高いと仮定し，動詞 A と名詞 B の単語間の概念的類似度を確認した結果，すべての単語分散表現で差があることを確認した．

今後は，類似度を含めた特徴量について検討し，機械学習に基づく手法により，「動詞 A ような名詞 B」タイプの直喩表現を自動抽出する手法を検討する．

表 2　各 50 文に対する正誤判定の結果

(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	9	7	5	4	1	6	11
誤	41	43	45	46	49	44	39
精度	0.180	0.140	0.100	0.080	0.020	0.120	0.220

(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	14	9	5	7	5	7	14
誤	36	41	45	43	45	43	36
精度	0.280	0.180	0.100	0.140	0.100	0.140	0.280

表 3　動詞 A と名詞 B 間の単語類似度

学習済みモデル	比喩である(除外文)	比喩でない(除外文)	差
①日本語Wikipedia エンティティベクトル	0.1821(5)	0.1698(2)	0.0123
②日本語 Wikipedia_fastText	0.2156(9)	0.2589(5)	0.0433
③HR領域の ワードベクトル	0.1291(55)	0.1729(37)	0.0438
④hottoSNS	0.1740(3)	0.2390(0)	0.0650
⑤朝日新聞 (cbow)	0.0926(4)	0.1098(0)	0.0172
(cbow-re)	0.1127(4)	0.1221(0)	0.0094
(glove)	0.1683(4)	0.2678(0)	0.0995
(glove-re)	**0.1962(4)**	**0.2972(0)**	**0.1010**
(skipgram)	0.2153(4)	0.2413(0)	0.0260
(skipgram-re)	0.3157(4)	0.3504(0)	0.0347
⑥chive	0.1787(0)	0.2291(0)	0.0504

参考文献

[1] 内海，“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”，JSAI2003 論文集，pp.1-4，2003.

[2] 田添他，“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] 宮脇他，“小説テキストに出現する直喩表現の抽出手法の検討”，IPSJ2022 講演論文集，pp.2-847-2-248，2022.

[4] 日本語 Wikipedia エンティティベクトル，https://github.com/singletongue/WikiEntVec

[5] 日本語 fastText モデル，https://github.com/Hironsan/awesome-embedding-models

[6] HR 領域のワードベクトル，株式会社ビズリーチ，https://www.bizreach.co.jp/technology/research/word2vec/

[7] 松野他，“日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築”，JSAI2019 論文集，pp.1-3，2019.

[8] 朝日新聞単語ベクトル，https://cl.asahi.com/api_data/wordembedding.html

[9] 真鍋他，“複数粒度の分割結果に基づく日本語単語分散表現”，NLP2019 発表論文集，pp.1407-1410，2019.