

コーパスとシソーラスを用いた比喩生成

Metaphor creation using corpus and thesaurus

佐藤 遼河^{*1}
Ryoga Sato

杉本 徹^{*2}
Toru Sugimoto

^{*1} 芝浦工業大学大学院 理工学研究科
Graduate School of Engineering and Science, Shibaura Institute of Technology

^{*2} 芝浦工業大学 工学部
College of Engineering, Shibaura Institute of Technology

We propose a method to create metaphors using corpus and thesaurus. Our system can create metaphors based on input tenors and grounds. We use word2vec to obtain word vectors for tenors and grounds, which are used to select vehicles considering cosine similarity. In order to output various vehicles that are congenial to the ground, we create a grounds-by-categories matrix using thesaurus. We evaluate appropriateness and usefulness of vehicles that are created by our system. As the result, it is shown that our system can output metaphors that are useful for user, though many of the metaphors created by the system are inappropriate. We compared vehicles created using Aozora Bunko corpus and NWJC. As the result, it is shown that our system can output vehicles that are congenial to user's literary style by using appropriate corpus.

1. 研究背景と目的

近年、小説投稿サイトの利用やセルフパブリッシングにより個人での執筆活動を公開する機会が増えている。それに伴い、読者にイメージが伝わりやすい表現を用いることへの需要が高まっている。イメージが伝わりやすい表現方法の一つとして比喩が挙げられる。しかし喩える語(喩辞)と喩えられる語(被喩辞)の関連性を考慮した上でことば選びをする必要があるため、比喩を作成することは難しい。

本研究は文章中のユーザが指定した語に対応する比喩を生成し、ユーザに提案するシステムを構築することが目的である。システムは入力された被喩辞と被喩辞の特徴を表す語(特徴語)から、執筆活動に用いることができる喩辞を複数生成し、提示する。システムが比喩の作成を支援することで、ユーザは容易にイメージが伝わりやすい表現を用いることができるようになる。比喩は直喩、隠喩、換喩、提喩に分類できるが本研究では直喩のみを扱う。

比喩の自動生成を目指した研究としては[北田 2001][中條 2017]などがある。北田らは被喩辞と喩辞の概念的な近さ、カテゴリの近さ、共起のしやすさを考慮するために EDR 電子辞書を用いて比喩の生成を行った。中條らは毎日新聞をコーパスとした共起行列を用いて比喩の生成を行った。

本研究では word2vec[Mikolov 2013]とシソーラスを用いて比喩を生成する。word2vec を用いることでユーザが執筆する文章のジャンルに合わせたコーパスを選ぶことができる。またシソーラスを用いることで特徴語と共起しやすい意味カテゴリに属する単語を優先的に出力する。

本稿では提案手法と青空文庫コーパスを用いて生成された比喩の評価実験の結果を報告し考察を行う。また別のコーパス

として国語研日本語ウェブコーパスを用いて比喩を生成し、青空文庫コーパスを用いて生成された比喩との比較を行った。

2. 提案手法

2.1 比喩生成の処理手順

提案手法の処理の流れを図1に示す。

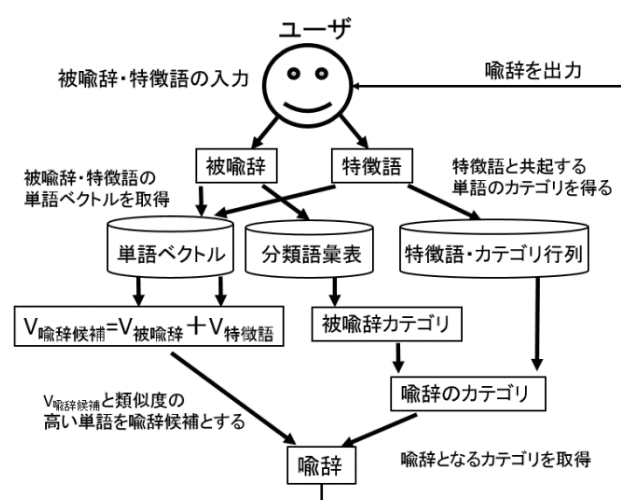


図1 提案手法の概要

システムの処理手順を以下に示す。

- ①ユーザが被喩辞と特徴語を入力
被喩辞は名詞、特徴語は形容詞、形容動詞、動詞のいずれかとする。例えば「少女はAのように美しい」という比喩を生成

連絡先: 佐藤遼河, 芝浦工業大学大学院 理工学研究科,
〒135-8548 東京都江東区豊洲 3-7-5,
E-mail: al14044@shibaura-it.ac.jp

- するのであれば、被喩辞として「少女」、特徴語として「美しい」を入力する。
- ②喩辞候補のベクトルを取得
- 被喩辞と特徴語に対応する単語ベクトルを取得する。この単語ベクトルは予め word2vec を用いて作成したものである。次に被喩辞のベクトルと特徴語のベクトルを加算して、得られたベクトルを喩辞候補のベクトルとする。喩辞候補のベクトルとコサイン類似度が高い単語ベクトルをもつ単語が喩辞候補となる。
- ③喩辞のカテゴリを取得
- システムは特徴語・カテゴリ行列を参照して、特徴語と共起しやすい単語のカテゴリを得る。ここから被喩辞のカテゴリを除いたものが喩辞のカテゴリとなる。
- ④喩辞の出力
- ②で取得した喩辞候補のうち、③で取得したカテゴリに属する語を喩辞として出力する。

2.2 特徴語・カテゴリ行列

特徴語を行、分類語彙表[国立国語研究所 2004]に基づく名詞のカテゴリを列とする共起行列を作成して用いる。

特徴語は名詞の特徴を表す形容詞、形容動詞、動詞の原形である。本研究では IPA 辞書において形容詞、名詞の形容動詞語幹、動詞の自立語に属する語であり、コーパス中に出現する語を扱う。

カテゴリは分類語彙表の中項目の名前である。分類語彙表の中項目の中で喩辞として適切でない語は予め取り除いた。喩辞として適切でない語とは抽象的な語、固有名詞、倫理的に不適切となりうる語である。

行列を作成するために、まずコーパス中の文章を係り受け解析し、特徴語と共起する名詞を含むカテゴリを抽出する。ここで、共起する名詞とは、特徴語が形容詞と形容動詞の場合は係り先、動詞の場合は係り元と係り先の文節中の名詞であるとする。例えば「冷たい雪に覆われた」という文の場合、「冷たい」と「覆われる」は「雪」と係り受けの関係にあるので、「冷たい」と「覆われる」は「雪」のカテゴリである「物質カテゴリ」と共起する。

本研究ではコーパスとして青空文庫に掲載されている作家 83 人分の文章を約 313MB 用いた。また係り受け解析には CaboCha を用いた。

行列の要素 a_{ij} の値は以下の式で求める。

$$a_{ij} = \frac{\text{特徴語}i \text{とカテゴリ}j \text{の単語が共起する回数}}{\text{カテゴリ}j \text{の単語の総出現回数}}$$

特徴語を表す行において要素の値が大きいものから順に喩辞のカテゴリの候補とする。

2.3 単語ベクトル

単語の分散表現ベクトルは word2vec の skip-gram を用いて 100 次元のベクトルを生成した。word2vec はコーパスからニューラルネットワークの学習を行い、単語の意味を表す固定長ベクトルを獲得する手法である。単語ベクトルはそれぞれの見出しの単語に対して float 型 100 次元の配列によって表現されている。入力された被喩辞、特徴語と一致する見出しを持つ単語ベクトルを求め、取得した 100 次元の配列をそれぞれ被喩辞ベクトルと特徴語ベクトルとする。

本研究ではコーパスとして前節で述べた青空文庫の文章において単語の活用形を基本形に直したものを用いた。表 1 に word2vec 実行時の各種パラメータを示す。

表 1 青空文庫をコーパスに用いたときの

word2vec 実行時のパラメータ		
CBOW or skip-gram	-cbow	0
次元数	-size	100
文脈長	-window	8
負サンプリング数	-negative	25
階層化 softmax	-hs	0
最低頻度閾値	-sample	1e-4
反復回数	-iter	15

3. 評価実験

3.1 生成された喩辞の適切性の評価

3.1.1 評価方法

生成された喩辞に対してアンケートによる定量的な評価を行った。10 種類の被喩辞と特徴語のペアをシステムに入力し、出力された喩辞を 5 人の被験者にすばらしい(4)、適切である(3)、判断が難しい(2)、不適切(1)の 4 段階で評価してもらった。アンケートでは被験者に『被喩辞』は〇〇のように『特徴語』という文を提示し、〇〇に喩辞を当てはめた場合の適切さを評価してもらった。例えば被喩辞「少女」、特徴語「美しい」の例では「少女は〇〇のように美しい」という文を提示する。

本実験ではシステムが生成する喩辞を喩辞候補ベクトルとのコサイン類似度が 0.5 以上かつ各カテゴリの上位 5 位以内の単語とし、最大で 50 語を出力した。

3.1.2 結果

表 2 に喩辞の出力結果の例を示す。

表 2 喩辞の出力例

入力 被喩辞, 特徴語	喩辞の カテゴリ	生成された喩辞
ひとみ, 光る	天地	二日月
		夜空
	物質	黒曜石
		夕焼け
ラジオ, 鳴る	資材	たいまつ
		イルミネーション
	物質	雷
		春雷
	道具	鐘
		ベル
人生, 儚い	言語	サイレン
		警鐘
	物質	うたかた
		泡沫
	心	夢幻
		追憶
	芸術	哀詩
		悲劇

次に喩辞の評価結果を表 3 に示す。

表 3 喩辞に対する評価の割合

喩辞数	4	3	2	1
477 個	4.2%	13.1%	12.3%	70.4%

生成された喩辞の 17.3% がすばらしい、または適切であるという結果が得られた。また入力した被喩辞と特徴語のペアのうち、コサイン類似度が 0.5 を上回る喩辞の候補が少なかったため、50 個の喩辞が得られないものが 3 組あった。

3.2 比喩の有用性の評価

3.2.1 評価方法

本実験では前節の実験で評価が高かった喩辞を被喩辞と特徴語のペアごとに上位 10 個ずつ用いて有用性の評価を行う。比較対象として文芸サークルに所属する大学生が作成した喩辞を被喩辞と特徴語のペアごとに 10 個ずつ用いる。6 人の被験者に被喩辞と特徴語を提示し、システムが生成した 10 個の喩辞と人が作成した 10 個、合計 20 個の喩辞をランダムに提示する。被験者は比喩として用いたいと思う表現を提示された喩辞の中から 5 個選択し回答する。

本実験で用いた喩辞の例を表 4 に示す。

表 4 評価に用いた喩辞の例

被喩辞：少女 特徴語：美しい	
システム	人
令嬢	サクラ
姫	女神
聖女	満月
騎士	初雪
女王	オーロラ
ヒロイン	星座
女優	花火
舞姫	宝石
花	夜空
人形	ダイヤモンドダスト

3.2.2 結果

被験者が選んだ喩辞のうちシステムが生成した喩辞の割合は 35.3%、人が作成した喩辞は 64.7% であった。このことからシステムが生成した喩辞に一定の有用性があることが分かった。被喩辞ごとの評価結果を図 2 に示す。

システムが生成した喩辞と人が作成した喩辞のうち、共通して得られたのは「被喩辞：ラジオ、特徴語：鳴る」に対する喩辞「雷」のみだった。この点から本システムは人が比較的思いつきにくい喩辞を提案できることが分かった。

生成された喩辞のうち 10 個中 9 個の比喩において有用な喩辞が得られた。有用な喩辞を得られなかった比喩は被喩辞「船」、特徴語「大きい」だった。これは「大きい」という特徴語が相対的な特徴語であるからだと考えられる。被喩辞「船」、特徴語「大きい」の評価に用いた喩辞とアンケートの結果を表 5 に示す。

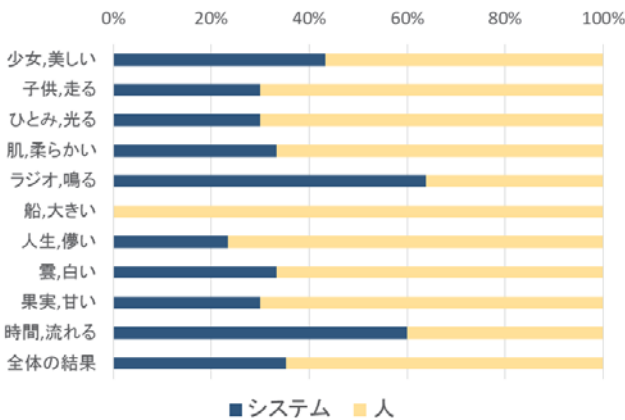


図 2 有用性の評価結果

表 5 「船が大きい」のアンケートの結果

被喩辞：船 特徴語：大きい			
システム		人	
太鼓腹	0人	山	5人
頭でっかち	0人	鯨	4人
巨体	0人	島	5人
海	0人	建物	1人
陸	0人	軍艦	3人
防波堤	0人	城	6人
ガスタンク	0人	要塞	3人
倉庫	0人	岩	2人
格納庫	0人	ドーム	1人
ロケット	0人	マンモス	0人

他の比喩における特徴語は「美しい」、「走る」、「光る」、「柔らかい」、「鳴る」、「儚い」、「白い」、「甘い」、「流れる」であり、いずれも絶対的な特徴である。例えば「花」の「美しい」という特徴は比較対象がある場合でも存在する。また「子猫」が「走る」という特徴も他の比較対象が「走る」場合でも失われない。一方「太鼓腹」の「大きい」という特徴は相対的な特徴であるため、「船」と比較した場合は失われる。つまり「太鼓腹」は「船」よりも小さい。「船」の「大きい」という特徴を強調するために「太鼓腹」という船より小さいものを用いたことで比喩として不適切になったと考えられる。

4. コーパスによる出力の差異

4.1 実験

コーパスによる出力の差異について調べる。本システムはユーザの執筆する文章のジャンルに合わせたコーパスを用いることで、より文章に適したことが選びができると期待できる。

本実験では word2vec を作成する際のコーパスとパラメータを変え、異なる単語ベクトルを用いて比喩を生成する。喩辞を生成する際に入力する被喩辞と特徴語は前節で用いた組み合わせと同じものを用いる。

4.2 使用する単語ベクトル

比較するベクトルは青空文庫をコーパスとして作成した単語ベクトル(以下「青空文庫単語ベクトル」と称する)と浅原らの作成した `nwjc2vec`[浅原 2017]である。`nwjc2vec` は国語研日本語ウェブコーパス(NWJC)から `word2vec` の CBOW を用いて生成した 200 次元の単語ベクトルである。国語研日本語ウェブコーパスはウェブ上の日本語テキストを利用した大規模コーパスである。

表 6 に `nwjc2vec` を生成する際の `word2vec` 実行時の各種パラメータを示す。

表 6 `nwjc2vec` を生成する際に用いられた

word2vec 実行時のパラメータ		
CBOW or skip-gram	-cbow	1
次元数	-size	200
文脈長	-window	8
負サンプリング数	-negative	25
階層化 softmax	-hs	0
最低頻度閾値	-sample	1e-4
反復回数	-iter	15

4.3 結果

はじめに前節の実験と同様にシステムが生成する喩辞をコサイン類似度が 0.5 以上かつ各カテゴリの上位 5 位以内の単語とし、最大で 50 語を出力した。その結果得られた喩辞は 108 個だった。青空文庫単語ベクトルが 477 個の喩辞を得たことと比較すると、得られた喩辞は少なかった。これはコサイン類似度が 0.5 以上の喩辞の候補が少なかったためである。そこで本実験では多くの喩辞を得るため、コサイン類似度が 0.4 以上の語を喩辞として出力する。

`nwjc2vec` を用いた場合の喩辞の出力例を表 7 に示す。

表 7 `nwjc2vec` を用いた喩辞の出力例

入力 被喩辞, 特徴語	喩辞の カテゴリ	生成された喩辞
ひとみ, 光る	植物	さゆり
		まゆみ
	物質	三日月
		流れ星
ラジオ, 鳴る	物質	陽子
		ルビー
	物質	遠雷
	道具	チャイム
	道具	ホイッスル
人生, 儚い	成員	DJ
	言語	サイレン
		警鐘
	社会	今生
		現世
	心	夢
		思い出
	芸術	物語
		悲喜劇

コサイン類似度が 0.4 以上かつ各カテゴリの上位 5 位以内の単語を最大で 50 語を出力するという条件で喩辞は合計 332 個得られた。

`nwjc2vec` を用いたことで青空文庫単語ベクトルとは異なる表現を得ることができた。例えば「被喩辞:ラジオ,特徴語:鳴る」に関して `nwjc2vec` を用いたことで「DJ」や「ホイッスル」などの喩辞を取得した。これは「DJ」や「ホイッスル」のような単語は青空文庫には用いられないが、ウェブでは一般的に用いられているからであると考えられる。一方、青空文庫単語ベクトルを用いた場合と同様に「遠雷」や「サイレン」や「警鐘」や「鐘」などの喩辞が得られた。`nwjc2vec` と比較すると、青空文庫単語ベクトルの方が同じ意味の語でも仰々しい言い回しの語が出力された。例えば `nwjc2vec` では「夢」や「思い出」が出力されているのに対し、青空文庫単語ベクトルでは「夢幻」や「追憶」が出力された。このことから `nwjc2vec` は現代が舞台の物語や説明的文章を書くときに用いる比喩に適している。一方青空文庫単語ベクトルは古風な表現が求められる歴史小説や伝奇小説、荘厳な世界観を表現するファンタジー小説を書くときに用いる比喩に適している。

`nwjc2vec` を用いた場合、多義語の問題があった。例えば「被喩辞:ひとみ,特徴語:光る」では被喩辞の「ひとみ」が `nwjc2vec` のベクトルでは眼球や目という意味だけでなく、女性の人名を表す語としての意味を表している。その結果コサイン類似度の上位に植物カテゴリからは「さゆり」や「まゆみ」、物質カテゴリから「陽子」などの女性の人名を表す語が出力されている。これはウェブで出現する人名の中に一般名詞としての意味をもつ語が多いからだと考えられる。

5. 結論

本研究では被喩辞と特徴語の入力に対してコーパスとシソーラスを用いて喩辞を生成する方法を提案した。今回行った実験により提案手法に一定の有用性があることが分かった。またコーパスを変えることで執筆する文章に合わせた喩辞が出力できることも分かった。提案手法を用いることでユーザは容易に比喩を用いることができるようになると期待できる。

今後の課題は出力された喩辞の適切性の低さの問題、相対的な特徴を表す喩辞をどのように扱うかという問題、多義語によって想定と異なるカテゴリの語が出力される問題の3つが挙げられる。

参考文献

[北田 2001] 北田純弥, 萩原将文: 電子辞書を用いた比喩による文章作成支援システム, 情報処理学会論文誌, Vol.42, No.5, pp.1232-1241 (2001)

[中條 2017] 中條寛也, 松吉俊, 内海彰: 意味空間に基づく文脈情報を用いた比喩生成, 情報処理学会研究報告, Vol.2017-NL-231, No.14, pp.1-10 (2017)

[国立国語研究所 2004] 国立国語研究所: 分類語彙表 増補改訂版, 大日本図書 (2004)

[Mikolov 2013] T. Mikolov, K. Chen, G. Corrado, and J. Dean: Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop (2013)

[浅原 2017] 浅原正幸, 岡照晃: `nwjc2vec`: 『国語研日本語ウェブコーパス』に基づく単語分散表現データ, 言語処理学会第 23 回年次大会発表論文集 (2017)