D-40

# 直喩文判定のための機械翻訳を用いた訓練データ獲得と分類モデルの考察

自見　仁太朗, 嶋田　和孝
九州工業大学知能情報工学科

## 1　はじめに

比喩の一種である直喩は，“ような” などの定型語 (喩詞) により比喩の対象を明示する表現である．しかし，喩詞として用いられる “ような” という語は，例示や婉曲の意味でも使用されるため，使われ方によって文意が大きく異なる．このような文を判別することは，文章を理解するうえで重要となる．

一般に，比喩検出を分類問題として機械学習で解くというやり方が考えられる．しかし，機械学習でこのタスクを行うためには直喩文と直喩表現を用いていない文 (リテラル文) それぞれの大量のテキストデータが訓練データとして必要になる．そして，これらのデータセットを人手で作成するには大きなコストがかかる．

そこで，本研究ではデータセットの自動獲得とそのデータセットを用いた直喩文判定モデルについて提案する．対象は “のような” と “のように” を含む文とし，直喩文判定を直喩文かリテラル文に分類する二値分類問題として定義する．また，データセットの自動獲得には機械翻訳を利用する．

機械翻訳を用いる目的は，分類対象の文の対訳文を獲得することである．近年ではニューラル機械翻訳の研究が進んでおり，機械翻訳による翻訳精度が大きく向上しているため，自動的に正確な対訳文が取得できることが期待できる [1]．正しい翻訳結果が得られていると仮定すれば，日本語文にはない特徴を得ることによって自動的に直喩文とリテラル文のデータセットを獲得することが可能になる．

## 2　データセット獲得手法

本節では，直喩文とリテラル文のデータセットを機械翻訳を用いて自動獲得する手法について述べる．まず，今回の手法では既存のデータセットである青空文庫と Wikipedia の本文データから “のような” と “のように” が含まれる文を対象として抽出する．

次に，抽出した文を機械翻訳にかけて対訳文を作成する．この際， python の googletrans ライブラリを用いる．

そして，作成した対訳文を参照しながら各文を直喩文とリテラル文に分けてそれぞれのデータセットを作成する．今回の分類対象の文に含まれる “のような” や “のように” といった語は，英文では比喩の意味合いで使われる場合は “like”，例示の意味合いで使われる場合は “as” と訳されるため，この二つの前置詞を特徴語として直喩文とリテラル文に分ける．このとき，どちらの条件も満たしていない文はデータセットから除いている．以下に直喩文とリテラル文それぞれのデータセットとして獲得される文の例を示す．

- 直喩文データセット - 対訳文に “like” を含む文
  **例:** まるで大きな子供のようなその無邪気さ。
  (It’s **like** a big child like that innocence.)
- リテラル文データセット - 対訳文に “as” を含む文
  **例:** このことは次のように明らかになるであろう。
  (This will be apparent **as** follows.)

表 1: データセット内訳

| | 直喩文 | リテラル文 |
|---|---|---|
| 青空文庫 | 26322 | 6889 |
| Wikipedia | 11053 | 18774 |
| 合計 | 37375 | 25663 |

表 2: それぞれのモデルによる分類結果 (F 値)

| | 機械翻訳のみ | NB | SVM |
|---|---|---|---|
| 直喩文 | 0.679 | **0.786** | 0.764 |
| リテラル文 | 0.498 | **0.673** | 0.628 |

前述の条件で分類を行った結果，集められた文は表 1 のような内訳となった．実験では，ここで集めたデータセットのうち直喩文とリテラル文をそれぞれ 20000 文ずつランダムに選出し，訓練データとする．

## 3　分類モデル

本研究では機械学習を用いて直喩文分類を行う．分類器には教師あり学習モデルの NB (Naive Bayes) と SVM (Support Vector Machine) を用いる．また，NB では Bag of Words を素性としたベクトル，SVM では Wikipedia のテキストデータで事前学習済みの word2vec モデルを用いて求めた文中の単語ベクトルの和を素性としたベクトルが入力となる．

## 4　実験

本実験では，提案手法の評価を行うために新たに直喩文とリテラル文それぞれの正解データセットを作成した．このデータセットは，2 節で獲得したデータセットからランダムに直喩文とリテラル文を 200 文ずつ主観で選出したものである．ただし，選出の際に多く見られた “以下のように” や “次のような” などの典型的な表現を含む文が重複してデータセットに含まれることは避けた．

機械学習の有効性を検証するため，機械翻訳のみで分類を行った場合と，2 節で獲得した訓練データで学習したモデルで分類した場合で精度を比較する．

これらの分類結果を表 2 に示す．表 2 を見ると，直喩文とリテラル文のデータセット双方に対して，機械学習モデルによる分類が機械翻訳のみのときよりも高い精度を示した．また，SVM より NB の方が高い精度を示した．この結果により，機械学習の有効性を示すことができた．

## 5　おわりに

本研究では，機械翻訳を用いて訓練データを集め，教師あり学習モデルによる直喩文判定に取り組んだ．今後は素性の見直しや訓練データの精査を行い，更なる精度の向上を目指す．

## 参考文献

[1] Yonghui Wu, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.