一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2021-36(2022-03)

疑似データ獲得による直喩文判定手法と分類モデルの考察

自見仁太朗† 　嶋田　和孝††

† 九州工業大学情報工学部知能情報工学科 〒820-8502 福岡県飯塚市川津 680-4
†† 九州工業大学 大学院情報工学研究院 知能情報工学研究系 〒820-8502 福岡県飯塚市川津 680-4
E-mail: †jimi.jintaro102@mail.kyutech.jp, ††shimada@ai.kyutech.ac.jp

あらまし　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方法が考えられる．しかし，機械学習でこのタスクを行うためには直喩文と非直喩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非直喩文に分類する二値分類問題として定義する．実験では獲得したデータセットを用いて複数のモデルで直喩文判定に取り組み，本タスクへの有効性を確認した．

キーワード　直喩文判定，疑似データ，訓練データ自動獲得，比喩

Simile identification based on machine learning using pseudo data acquisition

Jintaro JIMI† and Kazutaka SHIMADA††

† Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
†† Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
E-mail: †jimi.jintaro102@mail.kyutech.jp, ††shimada@ai.kyutech.ac.jp

Abstract　Simile is a kind of figurative language. It expresses the target of the figurative language by using comparators such as “like”. For understanding a sentence, it is important to distinguish whether the sentence is a simile or a literal. In this paper, we propose a pseudo dataset acquisition method for simile identification. We first constructed a dataset of simile and literal sentences using machine translation. Next, we define the simile identification task as a binary classification problem. We apply some machine learning approaches to the task. We show the validity of the pseudo dataset and the models in this task.

Key words　simile identification, pseudo data, automatic training data acquisition, figurative language

1. はじめに

自然言語処理の研究において，文章の意味や文脈の理解に関しては多くの課題が存在する．その中の一つが比喩理解である．一般に認知心理学の分野において，比喩理解には文字通りの情報だけでなく物事の持つイメージや概念などが認識過程に関わることが知られている．このような文章を機械が理解するには文字情報の他に概念知識や常識的知識が必要となる．したがって，比喩文を理解することは機械にとって困難なタスクといえ，文章理解においてそのような文を検知することは重要な課題といえる．

本研究では，比喩の中でも “ような” や “ごとく” のような定型語（喩詞）によって比喩の対象を明示する直喩に注目し，その中でも “のような” 又は “のように” を含む文を対象とする．直喩において喩詞として用いられる “ような” という語句は，例示や推測，引用といった様々な意味で用いられる．以下の例において (1) は直喩文であるが，(2) は非直喩文である．

- “ような” を含む文の例

（1）　天使の**ような**かわいい小鳥よ。（比喩）
（2）　具体的には次の**ような**問題である。(引用)

直喩文判定のためにはこのような文を判別する必要がある．

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方

法が考えられる．機械学習でこのタスクを行うためには直喩文と非直喩文それぞれの大量の訓練データが必要になるが，人手で作成するには大きなコストがかかる．

今回対象とした定型語は一般的な文章でも頻繁に登場し，既存の文章データの中にも多く存在する．そのため，それらを含む文に自動的にアノテーションをすることができれば大量の訓練データを得ることが可能になる．

本研究では，直喩文と非直喩文のデータセットを作成する方法として，対象となる文の対訳文を参照するという手法を提案する．このとき，対訳文は機械翻訳を用いて自動的に獲得する．日本語では比喩や例示など複数の意味に対して “ような” という語句が使われるが，英語に翻訳すると “ような” という語句はそれぞれの意味にしたがって “like” や “as” といった異なる前置詞で置き換えられることが考えられる．つまり，正確な対訳文が得られれば，英語の表現を利用して疑似的に直喩文か非直喩文かの判定を行うことができるようになる．こうして得られた疑似訓練データを基に機械学習による直喩文分類を行い，データセットやモデルの有効性を検証する．

2. 関 連 研 究

文における比喩性の検出の手法として，ルールベースで比喩の対象となる単語間（1 節の (1) だと「天使」と「小鳥」が比喩対象にあたる）の関係から類推する手法が存在する．田添ら [1] は「名詞 A のような名詞 B」という文型に限定してそれぞれの名詞の意味情報を用いたパターン分類による直喩文分類を試みた．しかし，田添らの手法では対象とする直喩文の形式が限定されていること，活用した知識ベースが小規模であることから，比喩検出を行う際に形式の不一致や未知語により判定を行えない文が出るため汎用性に乏しい．また，一般の文章にこの手法を適用する場合，比喩対象である名詞が判明している条件下でしか判定が行えないという点や，比喩対象の単語以外の情報が無視されるという点で問題が残る．近年では，ニューラルネットワークモデルを利用する研究が行われている．Gao ら [2] や Liu ら [3] は Bi-LSTM を用いた文中の比喩表現の検出を行っており，従来行われていたルールベースの手法よりも高い精度を記録している．しかし，この手法には大量の直喩文や非直喩文のデータセットが必要であり，それらのデータセットの人手での作成には大きなコストがかかることが課題となる．

データセット獲得の課題を解決する方法として，疑似データ獲得がある．疑似データ獲得は，事例の少ないデータの拡充や人手のアノテーションコスト削減を目的として行われる．特に画像処理の研究において，疑似データの獲得は画像の回転や反転などで比較的容易に行うことができ，それらのデータを使った学習による精度の改善も確認されている [4], [5]．しかし，言語処理の研究においては，対象の文の文章構造や文意を維持する必要があるため，単純に単語を置き換えたり語順を入れ替えたりして疑似データを増やすことは難しい．そこで，シソーラスや単語ベクトルの分散表現を用いて，できるだけ意味情報を保持したまま単語の置き換えを行うような手法がとられてきた [6], [7]．しかし，この方法では元となるアノテーション済みの少量のデータセットが必要となり，拡充できるデータ数も元のデータセットの量に依存する．また，拡充される文章は原文中の一部の単語が置き換わるだけであるため，真に多様な文章が得られているとは言い難い．

そこで，本研究では，機械翻訳を用いて自動的に直喩文と非直喩文の疑似訓練データを獲得する手法を提案する．近年では，ニューラル機械翻訳の研究 [8], [9] が進んでおり，機械翻訳の精度が大きく向上している．そのため，対訳文を参照することで本文の文意を加味した上で日本語にはない情報を得ることができ，その情報を元に擬似的なアノテーションを行うことが可能になる．また，この手法ではデータセットの作成に元となるアノテーション済みデータセットは必要なく，既存の大規模な文章データを活用することができる．このようにして獲得した疑似訓練データセットを用いて，複数の素性やモデルで直喩文判定の分類実験を行い，疑似訓練データや各分類器の本タスクへの有効性を検証する．

3. データセット

本節では，本研究で用いるデータセットについて説明を行う．3.1 節では，疑似訓練データ及びその獲得手法について説明する．3.2 節では，モデル評価のためのデータセットとその構築に伴って行ったアノテーション作業について説明をし，その結果について考察する．

3.1 疑似訓練データ

本研究では，機械翻訳を用いて “のような” 又は “のように” を含む文への自動的なアノテーションを行い，疑似訓練データを獲得する．3.1.1 節では，使用する翻訳器について説明する．3.1.2 節では，その翻訳器を用いてどのように疑似訓練データを獲得するかを述べる．

3.1.1 翻 訳 器

本研究では，mBART [8] というニューラル機械翻訳モデルを使用する．このモデルは，複数の言語でノイズ除去の事前学習が行われたシーケンス間モデルで構成されている．mBART にはいくつかのモデルが公開されており，本研究では 50 の多言語翻訳のために調整されたモデル[1]を使用する．

3.1.2 疑似訓練データ獲得手法

3.1.1 節で述べた翻訳器を用いたデータセット獲得の手順を述べる．まず，既存の文章データセットである青空文庫の本文データ[2]，Wikipedia の本文データ[3]，毎日新聞の 1995 年の記事データから対象とする “のような” 又は “のように” を含む文を抽出する．このとき，「このような（に），そのような（に），あのような（に），どのような（に）」といった “の” が指示語の一部であるパターンはデータセットから除外する．

次に，抽出した文を機械翻訳にかけて対象の文の対訳文を獲得する．その対訳文に対して次に示すルールを適用し，各文を直喩文と非直喩文に分けてそれぞれのデータセットを作成す

（注1）：https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
（注2）：https://github.com/aozorabunko/aozorabunko
（注3）：https://github.com/attardi/wikiextractor

表 1　疑似訓練データセット内訳

	疑似直喩文	疑似非直喩文
青空文庫	39987	14320
Wikipedia	14449	36337
毎日新聞	3939	1788
合計	58375	52445

る．今回の分類対象の文に含まれる “のような” や “のように” といった語句は，英文において比喩の意味合いで使われる場合は “like”，例示などの意味合いで使われる場合は “as” と訳される傾向があるため，この 2 つの前置詞を特徴語として直喩文と非直喩文に分ける．このとき，対訳文にどちらの特徴語も含まれていなかった場合，その文はデータセットから除外する．以下に直喩文と非直喩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．

- 疑似直喩文データセット - 対訳文に “like” を含む文

（3）　丁度鶏の脚**のような**骨と皮ばかりの腕である。

（3e）　It is just a bone and skin arm **like** a chicken’s leg.

- 疑似非直喩文データセット - 対訳文に “as” を含む文

（4）　このことは次**のように**明らかになるであろう。

（4e）　This will become clear **as** follows.

前述の条件で分類を行った結果，獲得した文は表 1 のような内訳となった．

3.2　評価データ

3.1 節で得たデータセットは直喩文と非直喩文への疑似的な分類であるため，データセット中にノイズが含まれている可能性がある．そこで，機械学習モデルによる直喩文判定における評価データを，人手のアノテーションによって構築する．ここで構築するデータセットは，3.1 節で得たデータセットより小規模なものであるが，人の主観評価によって分類されるため，よりノイズの少ないデータセットが構築できると考えられる．本節では，実施したアノテーションの設定について説明し，その結果を示す．

3.2.1　アノテーション設定

本研究では機械学習モデルによる直喩文分類の評価のため，人手評価のデータセットを作成した．そのデータセットを作るに当たって，9 人のアノテータにアノテーションを依頼した．

アノテーション対象の文は 3.1 節で収集した文から選出する．青空文庫，Wikipedia，毎日新聞の各ドメインごとに 300 文ずつランダムに選出し，計 900 文に対してアノテーションを依頼する．また，ここで選出した 900 文は 3.1 節で構築した疑似訓練データに含まれないように除外する．

各アノテータは各ドメイン 100 文ずつの全 300 文のアノテーションを行う．アノテータは 1 つの文に対して，その文が直喩文（s），非直喩文 (l)，判定不能 (u) のいずれであるかを判断し，それらの 3 種類のタグのうち一つを必ず付与する．9 人のアノテータに作業を依頼し，1 つの文に対して必ず 3 人のアノテータによるアノテーションが行われるようにする．

3.2.2　評価データの構築及び考察

3.2.1 節で説明したアノテーションの結果を示す．まず, Fleiss’ κ 値 [10] を用いて計算した各ドメイン及び全体のアノテーションの一致度を表 2 に示す．本研究では，アノテーションの結果，直喩文 (s) タグか非直喩文 (l) タグで 3 人のアノテータが付与したタグが完全一致したものを評価データとして用いる．その内訳を表 3 に示す．

表 2　アノテーション一致度

ドメイン	κ 値
青空文庫	0.437
Wikipedia	0.376
毎日新聞	0.341
全体	0.426

表 3　評価データ内訳 (完全一致)

	直喩文	非直喩文
青空文庫	124	55
Wikipedia	38	134
毎日新聞	55	94
合計	217	283

一般に，κ 値は 0.4〜0.6 の値をとるとアノテーション結果が適度に一致しているといわれている．表 2 を見ると，全体の κ 値は 0.4 を超えているため，アノテーションの信頼度はある程度保証されている．しかし，ドメインによっては κ 値が 0.4 を下回っているものもある．また，表 3 を見ると，アノテーションを行った全 900 文のうち半分以下の文でしか 3 人のアノテータのタグ付けが一致していない．これらの結果より，直喩文を判別することは人間にとっても判断の難しいタスクであることがいえる．

また，表 3 の各ドメインの結果より，ドメインごとに直喩文及び非直喩文の出現の傾向が違うことが分かる．このことから，“のような（に）” といった定型語は青空文庫のような読み物としての文章では比喩の意味で多く使用され，Wikipedia や新聞の説明的な文章ではそれ以外の意味で多く使われる傾向があることがいえる．

4.　提案モデル

本研究では，直喩文判定タスクを対象の文が直喩文であるか非直喩文であるかの二値分類問題として扱う．この節では，今回のタスクに用いる 3 つの分類モデルについて説明する．

4.1　SVM

SVM は，教師あり学習により 2 クラス分類を行う機械学習手法である [11]．このモデルでは，文中の単語の分散表現を用いることで文の大まかな文意を推測して直喩文判定を行うことを目的とする．

本研究でモデルの入力とする文章ベクトルの作成方法について述べる．まず，文を形態素解析器の MeCab[4]を用いて分かち書きする．次に，分かち書きされた文中の各単語を Wikipedia のテキストデータで事前学習済みの word2vec で 300 次元のベクトルに変換する．そして，文中の各単語ベクトルの総和を文

（注4）：https://taku910.github.io/mecab/

中の単語数で割った平均ベクトルを求め，その文の素性として使用する．

4.2 Naive Bayes

Naive Bayes は，ベイズの定理を利用して得られた確率値をもとに推測を行う教師あり学習モデルである．このモデルでは，文中の特徴的な単語を捉えることで直喩文判定を行うことを目的とする．

本研究でモデルの入力とする文章ベクトルの作成方法について述べる．まず，文を形態素解析器の MeCab で分かち書きする．そして，文中の各単語の出現回数の総和を計算し，その値に基づき Bag-of-Words によって文をベクトル化する．

4.3 BERT

BERT [12] は大規模なテキストコーパスを用いて事前学習を行った汎用言語モデルであり，目的のタスクに応じて追加学習 (ファインチューニング) を行うことで，様々な自然言語処理タスクに適応させることができる．このモデルでは，事前学習によって得られる汎用的な知識を使って，文脈情報を考慮した直喩文判定を行うことを目的とする．

本研究では，東北大学で公開されているモデル[5]を用いる．このモデルは，日本語 Wikipedia コーパスを用いて事前学習がされており，トークンへの分割は MeCab と WordPiece を用いて行われている．

モデルの入力では，一つの文章をトークンに分割して入力する．その際に，トークン列の先頭に [CLS] トークン，末尾に [SEP] トークンが追加される．BERT の出力層の [CLS] トークンの値を参照し，その文が直喩文か非直喩文かの判定を行う．

5. 実　　験

本節では 3 節で獲得した疑似訓練データ，評価データ及び 4 節で説明した提案モデルを用いて実験を行う．5.1 節では，3.1 節で獲得した疑似訓練データを用いて各分類モデルの学習を行った場合の分類結果と考察を述べる．5.2 節では，5.1 節の考察を受けて疑似訓練データの改善を行い，再度分類実験を行う．5.3 節では，BERT において 3.2 節で作成したアノテーションデータを用いてファインチューニングを行ったモデルと，疑似訓練データを用いてファインチューニングを行ったモデルの比較を行う．

5.1 各分類モデルの比較

本節では 3.1 節で獲得した疑似訓練データと 4 節で説明した提案モデルを用いて直喩文判定を二値分類問題として解く．評価データには 3.2 節で作成したデータセットを用いる．

5.1.1 実 験 設 定

本研究では，ベースラインとして機械翻訳のみで分類した結果を用いる．これは，評価データを機械翻訳にかけたときに対訳文に “like” を含むものを直喩文，それ以外の文を非直喩文と分類する二値分類モデルと仮定したものである．

次に，4 節で述べた各分類器の設定を述べる．Naive Bayes と SVM では，3.1 節で獲得した疑似訓練データから直喩文と非直喩文をそれぞれ 30000 文ずつランダムに選出して学習を行う．このとき，訓練データを選出する過程を 5 回行い，その平均値を算出する．SVM のパラメータはコストパラメータを 0.01 とし，それ以外はデフォルトの値を使用する．また，Naive Bayse のパラメータはすべてデフォルトの値を使用する．

BERT では，3.1 節で獲得した疑似訓練データから直喩文と非直喩文それぞれ 30000 文ずつをファインチューニングのための訓練データとして，直喩文と非直喩文それぞれ 300 文ずつを検証データとしてランダムに選出を行う．このとき，各データを選出する過程を 5 回行い，その平均値を算出する．これらの文を BERT に入力する際，各文の最大トークン数は 50 とし，文のトークン数が 50 を超えた場合は超過した後方のトークンは切り捨てる．最適化関数には，Adadelta を用い，学習率は 0.001 とし，損失関数は BCEWithLogitsLoss を用いる．また，バッチサイズは 128 とする．最大エポック数は 40 とし，その中で検証データの loss が一番小さかったものをモデルとして保存する．

表 4　各モデルごとの分類結果

	直喩文			非直喩文		
	Precision	Recall	F-score	Precision	Recall	F-score
ベースライン	0.627	0.613	0.620	0.708	0.721	0.714
SVM	0.637	0.830	0.721	0.830	0.637	0.721
Naive Bayes	0.652	0.813	0.724	0.823	0.668	0.737
BERT	0.600	0.876	0.713	0.854	0.553	0.671

5.1.2 実験結果及び考察

本節では，行った実験の結果及び考察を示す．このとき，評価には 3.2 節で作成した評価データを用いた．まず，ベースライン及び提案モデルにおける分類結果を表 4 に示す．下線はそれぞれの分類結果の最大値を示す．

ベースラインと比較すると，直喩文と非直喩文の F-score の平均ですべてのモデルがベースラインよりも高い値を得ている．このことから，本タスクにおける機械学習モデルの有効性が示された．モデル全体の傾向として，直喩文の Recall が高く，非直喩文の Recall が低いことがわかる．この原因として，非直喩文は直喩文以外の全般の文を示すため，直喩文よりも定義範囲が広いことで非直喩文の特徴が分散し，直喩文よりも分類が難しかったと考えられる．また，3 つのモデルの中では Naive Bayes が最も高い平均 F-score を得た．このことから直喩文判定において，直喩文や非直喩文に含まれる特徴的な単語に注目することが重要といえる．

次に，3 つのドメインごとの結果を確認する．表 5 に表 4 の結果における各ドメインの Recall を示す．Recall を示すのは，各ドメインごとの判定の偏りの有無を確認するためである．

この結果を見ると，モデルやドメインごとに直喩文と非直喩文の Recall の差が顕著に表れていることがわかる．Naive Bayes と SVM では各ドメインでほぼ同様の傾向が見られ，青空文庫では直喩文，Wikipedia では非直喩文への判定の偏りが大きい．これは 3.1 節で獲得した疑似訓練データの各ドメインのデータ数の偏りと一致している．このことから，分類器が直喩文と非直喩文の特徴を見ておらず，青空文庫の文体であれば直喩文，

(注5)：https://github.com/cl-tohoku/bert-japanese

表 5　各ドメインの Recall

		SVM	Naive Bayes	BERT
青空文庫	直喩文	0.984	0.982	0.944
	非直喩文	0.109	0.145	0.211
Wikipedia	直喩文	0.552	0.447	0.658
	非直喩文	0.881	0.928	0.784
毎日新聞	直喩文	0.691	0.698	0.877
	非直喩文	0.596	0.598	0.423
全体	直喩文	0.830	0.810	0.876
	非直喩文	0.636	0.668	0.553

表 6　データ数の偏りを解消したデータセット内訳

	疑似直喩文	疑似非直喩文
青空文庫	14449	14320
Wikipedia	14449	14320
毎日新聞	3939	1788
合計	32837	30428

Wikipedia の文体であれば非直喩文というように，ドメインごとの特徴を検知して判定を行っている可能性がある．また，毎日新聞の評価データにおいては，直喩文や非直喩文の Recall の差は他の 2 つのドメインに比べて小さい．これは，毎日新聞のデータセットが青空文庫や Wikipedia のデータセットよりデータ数が少ないため，分類結果への影響が少なかったと考えられる．

BERT では，青空文庫においては他の 2 つのモデルと同様に直喩文への大きな判定の偏りが見られるが，Wikipedia では偏りは他の 2 つのモデルに比べて小さい．これは，BERT では事前学習に Wikipedia のデータセットを用いているため，その知識がファインチューニング時のデータ数の偏りによる影響を抑えたと考えられる．

5.2　データ数の偏りを解消したデータによる学習

本節では，5.1 節の結果を受け，疑似訓練データにおけるドメインごとのデータ数の偏りを解消した上で，再度実験を行う．

5.2.1　実 験 設 定

今回の実験では，表 1 のデータセットを元にして疑似訓練データセットを再構築する．表 1 のデータセットにおいて，青空文庫の文が直喩文，Wikipedia の文が非直喩文のデータの割合を多く占めている．そこで，青空文庫の直喩文を Wikipedia の直喩文のデータ数，Wikipedia の非直喩文を青空文庫の非直喩文のデータ数に揃えるようにデータセットからランダムに削除することで，各データセットにおけるドメインごとのデータ数の偏りを解消する．毎日新聞のデータは他の 2 つのドメインと比べてデータ数が少ないため，そのまま用いる．新たに作成したデータセットを表 6 に示す．表 1 のデータセットからの変更点を下線で示す．

また，使用する疑似訓練データセット以外の設定は 5.1 節で述べた実験設定と同様の設定で実験を行った．

5.2.2　実験結果及び考察

表 7 に各モデルごとの分類結果を示す．下線は各分類結果の最大値を示す．また，表 8 は 5.1.2 節と同様にモデルの各ドメインごとの Recall を示し，議論していく．

表 7　各モデルごとの分類結果（偏りを解消したデータセット）

	直喩文			非直喩文		
	Precision	Recall	F-score	Precision	Recall	F-score
ベースライン	0.627	0.613	0.620	0.708	0.721	0.714
SVM	0.657	0.768	0.708	0.795	0.692	0.740
Naive Bayes	0.657	0.784	0.715	0.806	0.685	0.740
BERT	0.569	0.838	0.678	0.806	0.514	0.628

表 8　各ドメインの Recall（偏りを解消したデータセット）

		SVM	Naive Bayes	BERT
青空文庫	直喩文	0.839	0.797	0.813
	非直喩文	0.455	0.527	0.465
Wikipedia	直喩文	0.658	0.700	0.721
	非直喩文	0.810	0.815	0.679
毎日新聞	直喩文	0.698	0.829	0.982
	非直喩文	0.660	0.589	0.304
全体	直喩文	0.768	0.784	0.838
	非直喩文	0.692	0.685	0.514

表 5 と表 8 を比較するとほとんどのドメインにおいて直喩文と非直喩文の Recall の差が小さくなっていることがわかる．このことから，訓練データにおける各ドメインのデータ数の偏りがドメインごとの判定結果に影響を及ぼしていることがいえる．

また，表 4 と表 7 の結果を比較すると，すべてのモデルにおいてデータセットを変更したあとの方が直喩文と非直喩文の平均 F-score はやや減少する結果となった．これは，各ドメインの多数派に寄せることで正しく判定できていたものが誤判定されるようになったことが原因だと考えられる．しかし，Naive Bayes と SVM においては，依然としてベースラインよりも高い平均 F-score を記録しており，本タスクにおけるモデルの有効性が示された．

5.3　アノテーションデータでの BERT のファインチューニング

前節までの実験で使用していた 3.1 節で獲得した訓練データは機械翻訳によって疑似的なアノテーションを行っているため，データ中にノイズが存在し，判定結果に悪影響を与えている可能性がある．また，一般に BERT では事前学習の知識を利用することで，少量のデータでのファインチューニングでも高い精度を発揮することが知られている．そこで，3.2 節で人手のアノテーションによって獲得したノイズの少ないデータを BERT のファインチューニングに用いて実験を行い，5.2 節での BERT の実験結果との比較を行う．

5.3.1　実 験 設 定

今回の実験では，3.2 節で獲得した評価データを訓練データ：検証データ：テストデータ = 8：1：1 に分割し，10 分割交差検証を行う．BERT の設定はバッチサイズを 32 とし，その他は 5.1.1 節で述べた設定と同じとする．

5.3.2　実験結果及び考察

表 9 に実験結果を示す．疑似訓練データを使用した際の実験結果は表 7 の結果と同様のものである．評価データを用いた際

表 9　ファインチューニングをするデータセットの違いによる分類結果の比較

	直喩文			非直喩文		
	Precision	Recall	F-score	Precision	Recall	F-score
疑似訓練データ	0.569	0.838	0.678	0.806	0.514	0.628
評価データ	0.792	0.223	0.319	0.608	0.919	0.726

の実験結果は前述の実験設定で述べた 10 分割交差検証を行った際の平均値を示す．

表 9 を見ると，評価データを使用したモデルでは直喩文の Recall と F-score が著しく低い．そして，評価データが異なるため一概に比較を行うことはできないが，疑似訓練データを使用したモデルの方が直喩文と非直喩文の平均 F-score は高い．

また，評価データを使用したモデルによる 10 分割交差検証で，F-score の標準偏差は直喩文では 0.080，非直喩文では 0.054 だった．疑似訓練データを使用したモデルの 5 回の実験では，F-score の標準偏差は直喩文と非直喩文ともに 0.004 程度だった．このことから，評価データを使用した場合は疑似訓練データを使用した場合に比べて，学習データやテストデータの組み合わせによって判定のばらつきが大きいことがわかる．

これらの結果から，直喩文分類タスクにおいての BERT でのファインチューニングに関して，3.2 節で収集した直喩文と非直喩文それぞれ 200 件程度のデータでは不十分であることがいえ，ノイズが多少含まれていても大規模な訓練データを構築することの有効性を示すことができた．

6. おわりに

本研究では，“のような” 又は “のように” を含む文を対象として機械学習による直喩文判定を行った．また，このタスクで使用する訓練データを機械翻訳を用いて獲得する手法について提案した．具体的には，大規模な既存文章データセットから取得してきた対象の文に対して，機械翻訳で得た対訳文を参照することで擬似的なアノテーションを行い，直喩文及び非直喩文それぞれのデータセットを作成した．

実験の結果，3 つの提案モデルすべてでベースラインである機械翻訳のみで分類した結果よりも高い平均 F-score を得ることができ，本タスクにおける機械学習の有効性が示された．また，提案モデルの中でも Naive Bayes モデルが一番良い結果を得た．次に，各ドメインの結果を比較したところ，直喩文と非直喩文の Recall に偏りが見られた．これらの偏りは直喩文と非直喩文データセット内の各ドメインのデータ数の偏りと一致している．このことから，直喩文や非直喩文の特徴を捉えて分類しているのではなく，ドメインごとの特徴を捉えて分類を行っている可能性が考えられる．

そこで，集めた直喩文と非直喩文データセットの各ドメインのデータ数の偏りをなくして，再度実験を行った．実験の結果，各ドメインの直喩文と非直喩文の Recall の偏りが緩和された．このことから，本手法で複数ドメインからデータセットを獲得する場合，データセットを占める各ドメインの割合に注意する必要があることが示された．また，各モデルの平均 F-score は前回の実験よりも下がったが，Naive Bayes と SVM においてはベースラインよりも高い値を保っており，本タスクにおける有効性が変わらず示された．

最後に，BERT のファインチューニングにおける本研究で獲得した疑似訓練データセットの有効性について検証した．実験結果より，小規模なデータセットでは本タスクへの学習が十分とはいえず，疑似訓練データのような大規模なデータセットでファインチューニングを行うことの有効性が示された．

今後の課題として，疑似訓練データの精査や素性の見直しが挙げられる．今回，疑似訓練データのアノテーションには “like” や “as” という語句を判断基準においたが，より適切な判断基準を追加することができればアノテーションの精度を上げることができる．また，今回のモデルへの入力では文中の単語を一様な重みで扱ったが，直喩文にとって重要な単語に重み付けができれば更なる精度の改善が見込める．これらに取り組み，直喩文判定タスクの更なる精度向上を目指したい．

文　献

[1] 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, Vol. 10, No. 2, pp. 43–58, 2003.

[2] Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 607–613. Association for Computational Linguistics, 2018.

[3] Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1543–1553, 2018.

[4] 党紀, 松山当也, 全邦釘, 史紀元, 松永昭吾. UAV 画像における損傷自動認識のための深層学習と精度向上手法に関する検討. AI・データサイエンス論文集, Vol. 1, No. J1, pp. 596–605, 2020.

[5] 小林賢一, 辻順平, 能登正人. ディープラーニングを用いた画像処理による農作物病害診断への data augmentation の応用. 情報処理学会 第 79 回全国大会講演論文集, Vol. 79, No. 2, pp. 289–290, 2017.

[6] 西本慎之介, 能地宏, 松本裕治. データ拡張による感情分析のアスペクト推定. 言語処理学会第 23 回年次大会, 発表論文集, pp. 581–584, 2017.

[7] 颯々野学. サポートベクタマシンを使った文書分類における仮想事例の利用. 自然言語処理, Vol. 13, No. 3, pp. 21–35, 2006.

[8] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 726–742, 2020.

[9] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[10] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, p. 378, 1971.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.