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
NLC2023-26(2024-03)

複数の訓練データを用いた直喩文判別

自見仁太朗† 　嶋田　和孝††

† 九州工業大学大学院情報創成工学専攻 〒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) によって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，使い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理解に於いて重要である．一般に，比喩判別を二値分類問題として機械学習で解くことが考えられる．しかし，機械学習モデルを学習させるには大量のラベル付きデータ（直喩 or 非直喩）が必要となる．このようなデータセットを新たに構築するには大きなコストがかかる．そこで，本研究では少量のラベル付きデータに加えて，疑似ラベリングデータ，他言語データを用いて学習を行い，これらのデータセット及びその組み合わせの有効性を確認した．

キーワード　直喩文判別，疑似データ，多言語学習，比喩

Simile identification using several types of training data

Jintaro JIMI† and Kazutaka SHIMADA††

† , 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 use several types dataset for simile identification task. The several datasets are small human annotated datasets, pseudo-labeling dataset, and other language datasets. We show the validity of the combination of several data in this task.

Key words　simile identification, pseudo data, multilingal training, figurative language

1. はじめに

自然言語処理分野の研究において，文章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表現の理解である．一般に認知言語学の分野において，比喩を理解するためには文字通りの情報だけではなく物事の持つイメージや概念などの情報が認識過程において重要であることが知られている．このような表現を機械が理解するためには文字情報だけでなく，物事の概念知識や常識的知識が必要となる．そのため，比喩表現を理解することは機械にとって困難なタスクであり，文章理解においてそのような文を検知することが重要な課題であるといえる．

本研究では，比喩の中でも「ような」や「ごとく」といった定型語 (喩詞) によって比喩の対象を明示する直喩表現に注目し，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，文中で例示や婉曲といった喩詞以外の役割を持つ場合も多い．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．直喩文判別では下記のような文を見分ける必要がある．

- 「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)　大谷翔平の**ような**野球選手になりたい。(例示)

一般に，このようなタスク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，機械学習で分類問題に取り組む場合，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．日本語におけるこのような大規模データセットは不足しており，新たに人手で構築するのは大きなコストがかかる．

本研究では，このリソースの不足を補うために，少量の直喩

判別アノテーションデータに加えて，疑似ラベリングデータセット，他言語のデータセットを用い，これらのデータセットやその組み合わせの有効性を確認する．

2. 関 連 研 究

直喩文判別の手法として，比喩の対象である喩詞と被喩詞 (前節の (1) では「天使」が喩詞で「小鳥」が被喩詞である) の関係を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各名詞の意味情報を知識ベースから獲得し，パターン分類による直喩文判別を試みた．しかし，田添らの手法は対象とされる直喩文の形式が限定される点や，知識ベースに存在しない名詞に関して類推が行えない点などで汎用性に欠ける．また，この手法を適用する場合は文章中の喩辞と被喩辞が明確になっている必要があるが，喩辞と被喩辞を文章から抽出することは困難なタスクの一つである [2].

近年では，ニューラルネットワークモデルを用いた研究が行われている．Gao ら [2] や Liu ら [3] は Bi-LSTM を用いて文中の比喩表現の検出に取り組んでおり，従来のルールベース手法よりも高精度を記録している．しかし，このようなモデルを学習するには正解データとなる大量のラベル付きデータが必要となる．近年では大規模な事前学習モデル BERT [4] の登場により，十分な学習に必要なデータ量は減少したが，依然としてより良いモデルを構築するためには一定量の高品質なデータセットの構築が必要であり，その作成コストは課題となる．

このような背景から，データセットの不足を補うために様々な研究が行われてきた．例えば，疑似データの作成・取得は，事例の少ないデータの拡充や人手アノテーションのコスト削減のために一般的に取られる手法の一つである．特に，画像処理の研究において疑似データによるデータ拡充は画像の反転や回転によって比較的容易に行うことができる [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崩してしまうため，単純に単語の置換や削除，挿入といった処理でデータ拡充を行うことは難しい．

そこで，シソーラスや単語の分散表現を用いて，できるだけ元の文章の文意を保持したまま単語を置換するような手法がとられてきた [6] [7]．しかし，この手法で得られる疑似データは元の文章に大きく依存するため，真に多様な文章が得られるとは言い難い．また，置き換えられる単語には限りがあるため，得られるデータ量は元のデータ量に大きく依存する．そのため，本研究では私たちの先行研究 [8] に倣い，機械翻訳を用いて疑似データの獲得を行う．近年では，ニューラル機械翻訳 [9] の研究が進んでおり，機械翻訳の精度が大きく向上している．そこで，機械翻訳で得た対訳文の情報を加味することで疑似的なラベリングを行うことを可能とした．

データセットの不足を補う他の手法として，マルチタスク学習がある．マルチタスク学習とは，異なるタスクについて学習した複数のモデル間で相互学習を行い，それぞれのタスクで得られた知識を他のモデルの学習に取り入れる手法である [10]. Zhu ら [11] は疾患分類タスクにおいてマルチタスク学習の枠組みを取り入れ，同一タスクを複数言語のモデルで相互学習させることで当該タスクの精度向上を確認した．本研究では，このマルチリンガルでの相互学習手法を直喩判別タスクに適用する．

表 1　データセット内訳

	言語	直喩	非直喩
疑似ラベリングデータ	日本語	106968	57501
Human Annotation 1	日本語	217	283
Human Annotation 2	日本語	300	723
Multilingal Simile Dialogue Dataset	英語	5515	5904
	中国語	3576	4570
Chinese Simile Recognition Dataset	中国語	5088	6249

3. データセット

本節では，実験で用いるデータセットについて述べる．3.1 節では疑似ラベリングデータ及びその獲得手法について説明する．3.2 節では人手アノテーションデータの構築手順とその結果を示す．3.3 節では，5. 節で用いる他言語直喩判別データに関して説明する．各節で説明するデータセットの内訳について表 1 に示す.

3.1 疑似ラベリングデータセット

本研究では私たちの先行研究 [8] に倣い，機械翻訳を用いて得られた他言語の情報を基に「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，疑似ラベリングデータを獲得する．3.1.1 節では使用する翻訳器について述べる．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．獲得したデータセットの内訳は表 1 に示す.

3.1.1 翻　訳　器

本研究では，mBART [12] という機械翻訳モデルを使用する．このモデルは，複数の言語で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．このモデルは 50 の言語において任意のターゲット言語 ID を与えることでペア間での直接翻訳を行うことができる．本研究では日英翻訳のみを利用する.

3.1.2 疑似ラベリングデータ獲得手法

本節では 3.1.1 節で述べた翻訳器を用いた疑似ラベリングデータセット獲得の手順を説明する．まず，既存の日本語文章コーパス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 年の記事データを対象として文章の抽出を行う．抽出対象は「のような」又は「のように」を含む文とする．対象の文に対して上記の文字列のマッチングを行い，各コーパスから抽出する．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ストップワードとして設定し，データセットから除外している.

次に抽出した文を機械翻訳にかけて対訳文 (英文) を獲得す

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

る．その対訳文に対して次に示すルールを適用し，各文に直喩・非直喩のラベルをそれぞれ付与した．

- 疑似直喩 - 対訳文に “like” を含む文

(4) 丁度鶏の脚**のような**骨と皮ばかりの腕である。

(4e) It is just a bone and skin arm **like** a chicken’s leg.

- 疑似非直喩 - 疑似直喩ではない文

直喩には特徴語 (Comparater) が存在し，日本語ではその一つとして今回対象としている「ような (に)」が挙げられる．同じように他の言語でも直喩を表す特徴語が存在する場合がある．今回用いる英語では一般的な特徴語の一つとして前置詞 “like” が挙げられる．そこで，獲得した対訳文に “like” が含まれている文を疑似直喩とし，それ以外の “like” が含まれていない文を疑似非直喩とするように対象の文にラベルを付与した．

また，該当するデータを抽出したのち，疑似直喩データと疑似非直喩データそれぞれに含まれる青空文庫と Wikipedia データセットの文数を均等にするようにダウンサンプリングを行った．毎日新聞のデータセットに対してこの処理を行わないのは，当データセットが他のデータセットに比べて文数が極端に少ないためである．

3.2 人手アノテーションデータセット (日本語)

本節では，2 セットの人手アノテーションデータの構成及び作成手順について説明する．一つは，私たちの先行研究 [8] で構築したデータセットである．もう一つは，先行研究の結果を踏まえ，今回新たに作成したデータセットである．以降では，先行研究のデータセットを HA1 (Human Annotation 1)，今回構築したデータセットを HA2 (Human Annotation 2) とする．3.2.1 節でそれぞれのアノテーションの設定について説明し，3.2.2 節でその結果について示す．また，これらのデータセットは 3.1 節の疑似ラベリングデータセットと同じドメインからデータを取得しているが，それぞれのデータセットは独立である．表 1 にこれらのデータセットの内訳を示す．

3.2.1 アノテーション設定

HA1 は先行研究 [8] で構築したデータセットである．このデータセットを構築するに当たって，青空文庫，Wikipedia，毎日新聞の各ドメインごとに「のような (に)」を含む文を 300 文ずつをランダムに選出した．そして，これら計 900 文に対して 9 人のアノテータにアノテーションを依頼した．

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．アノテータはそれぞれの文に対して，直喩文，非直喩文，判定不能のいずれか 3 値のラベルを付与する．このとき，1 つの文に対して必ず 3 人のアノテータによってアノテーションが行われるようにした．

次に，HA2 のアノテーション設定について説明する．このデータセットを作成するに当たって，3.1 節で作成した疑似直喩，疑似非直喩データの各文から各 800 文をランダムに選出した．そして，これら計 1600 文に対して 12 人のアノテータにアノテーションを依頼した．

各アノテータはこのうちランダムな 400 文に対してアノテーションを行う．今回のアノテーションでは表 2 に示す基準で 0〜4 の値をそれぞれの文に付与する．確信度は，少しでもラベル付与の判断に迷った場合に確信度低とするように指示をした．

表 2 HA2 で付与したラベル

ラベル	ラベルの意味
4	直喩 (確信度高)
3	直喩 (確信度低)
2	非直喩 (確信度低)
1	非直喩 (確信度高)
0	判別不能 (特殊タグ)

また，今回のアノテーションでは 0 ラベルを特殊タグと位置づけ，ラベルの付与に制限を与えた．このラベルは，与えられた文が非文であったり，前後文脈が不足していたりして直喩か非直喩か判断できない場合に付与される．このラベルを付与できる最大数を各アノテータ 8 件とし，9 件目以降は判断が難しい場合でもそれ以外のラベルを付与するように指示した．

3.2.2 アノテーション結果

まず，アノテーションの一致度について Fleiss’ κ [13] を用いて評価を行う．この値を計算する際に，HA2 に関しては 4・3 ラベル及び 2・1 ラベルを同一ラベルと見做して計算を行っている．計算の結果，それぞれのデータの κ 値は HA1 が 0.426，HA2 が 0.488 となった．一般に κ 値は 0.4〜0.6 の値であればアノテーション結果が適度に一致しているといわれる．HA1，HA2 双方のデータはその条件を満たすため，これらのアノテーションデータの信頼度はある程度保証されているといえる．

本実験では，3 人のアノテーションが完全一致した場合のデータのみを用いてデータセットを構築する．

3.3 他言語データセット

この節では 5. 節で用いる他言語直喩判別データに関して説明する．本研究では 2 種類の他言語データセットを用いる．表 1 にこれらのデータセットの内訳を示す．

1 つ目は，Multilingal Simile Dialogue Dataset [14](MSD) である．このデータセットは英語と中国語のダイアログコーパス (Reddit-dialogue，LCCC，PchatbotW) から抽出した文章から構築されている．抽出対象は機械翻訳した際に英語，中国語の双方で特徴語 (“Like”，「像」など) を含む一定長のダイアログとし，それに対して直喩か非直喩かのラベルを人手で付与した．このとき，それぞれの文に最低 3 人のアノテータがラベルを付与し，その Fliess’ κ 値は 0.61 であった．

2 つ目は，Chinese Simile Recognition Dataset [3](CSR) である．このデータセットは中国の学生が書いたエッセイのうち，特徴語である「像」が含まれる文を抽出し，アノテーション対象としている．そして，それらの文に対して 2 人のアノテータが直喩か非直喩かのラベルを付与した．

以下の章では各データセットを MSD_{en}，MSD_{zh}，CSR_{zh} と記述する．

4. 日本語データを用いた学習

本節では疑似ラベリングデータと人手アノテーションデータを用いてモデルの学習を行った結果を示す．4.1 節では，本実験で分類モデルとして用いる事前学習モデル BERT について説

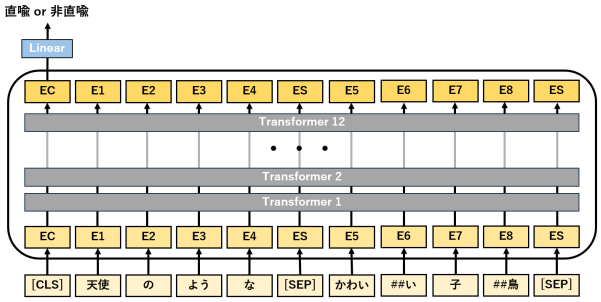

図 1　BERT の概略

明する．4.2 節で実験の設定について説明し，4.3 節で実験結果について示す．

4.1　BERT

BERT [4] 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．本タスクにおけるモデルの概要を図 1 に示す．

BERT に入力される文は BertJapaneseTokenizer によってサブワード単位で分割されて入力される．サブワード分割されている単語は図 1 中の##で示されている．また，図 1 中の [CLS] と [SEP] は BERT における特殊トークンの一種である．[CLS] トークンは後ろに続く文章全体の特徴を示すトークンであり，今回のような文章分類タスクにおいては最終的な特徴量として用いられる．[SEP] トークンは主に文章の終端や発話の区切りを示すために使用されるトークンである．BERT に入力された各文は 12 層の Transformer 層を通り，各トークンが 768 次元の出力ベクトルを得る．最後に [CLS] トークンを Linear 層を用いて 2 次元に圧縮し，2 値分類の結果として扱う．

本実験では，東北大学で公開されているモデル[4]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行われている．モデルに対して，疑似ラベリングデータ，人手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い，直喩判別タスクに適応させる．

4.2　実 験 設 定

本実験では 3.1 節の疑似ラベリングデータ及び 3.2 節の人手アノテーションデータを用いて BERT モデルを学習する．それぞれのデータセットを訓練に用いるときに直喩データ，非直喩データの数を合わせるため，ダウンサンプリングを行う．このときのデータの選出はランダムに行う．また，この選出されたデータを更に 9 : 1 に分割し，訓練データと開発データとする．

各文章は図 1 のように特徴語「ような (に)」の後ろに [SEP] トークンを挿入し，BERT に入力する．これは，文章の喩辞が含まれる部分と被喩辞が含まれる部分を明示的にモデルに示す意図がある．

本実験の評価データは人手アノテーションデータのうち，学習に使われていないものとする．つまり，HA1 で学習されたモデルの評価を HA2 で行う．同じように HA2 で学習されたモデルの評価は HA1 で行われる．疑似ラベリングデータのみで学習されたモデルに対しては，HA1，HA2 の双方で評価を行う．

次に，本実験で用いる BERT のハイパーパラメータ設定を説明する．まず，最適化関数に Adadelta，損失関数に BCELoss を用いる．また，学習率を 1e-3，重み減衰を 1e-5 に設定する．入力の最大トークン長は 64 とし，バッチサイズは疑似データで学習するときは 128，HA1 及び HA2 で学習するときは 16 とした．最大エポック数は 60 として Early Stopping を導入し，10 エポック間モデルの更新がなかった場合，学習を打ち切る．

本実験のベースラインとして, 疑似ラベリングデータ (pseudo), 2 種の人手アノテーションデータ (HA1，HA2) をそれぞれ単体で学習したモデルを用いる．このとき，訓練データを選定してモデルを学習する過程を 3 回行い，その平均値を算出する．

また，疑似ラベリングデータと人手アノテーションデータを組み合わせた 2 つの学習方法を提案し，比較する．一つ目は，人手アノテーションデータのデータ数を疑似データによって拡充する方法である．この方法では，それぞれの人手アノテーションデータに疑似直喩文，疑似非直喩文のデータを一定数拡充したデータセットを用いて学習を行う．今回は 100，10000 件の 2 パターンを比較する．二つ目は，疑似ラベリングデータでファインチューニングされたモデルに対して，人手アノテーションデータで再度ファインチューニングを行う手法である．以上の 2 つの手法に関してもモデルの学習過程を 3 回行い，その平均値を算出する．

4.3　実 験 結 果

表 3 および表 4 に各実験結果を示す．表の値はそれぞれ直喩，非直喩の F 値とその平均 F 値である．また，表中の太字は最高 F 値を示し，下線は 2 番目に高い F 値を示す．

まず，単体で学習した結果 (表中の「疑似」と「人手」) について比較する．それぞれの表から，疑似データで学習するよりも人手データで学習する方がより高い F 値を示すことがいえる．これは疑似データがルールベースのラベリングであるため，ノイズを多量に含むことが原因の一つであると考えられる．

次に，疑似ラベリングデータと人手アノテーションデータを組み合わせた手法について考察する．表中の HA1 + N，HA2 + N はそれぞれの人手データに直喩，非直喩それぞれ N 件の疑似データを拡充した結果を示す．これらの結果は，人手アノテーションデータのみで学習したときよりも F 値の低下が見られた．また，拡充したデータ数がより多い +10000 の方が F 値の低下する幅はより大きい．この結果より，疑似データの拡充はノイズとなり，学習への悪影響を与えることがいえる．

一方で，追加学習の手法では，人手アノテーションデータのみの学習よりも F 値の向上が見られた．本手法の特徴として，大規模な疑似ラベリングデータでおおまかな学習をしたのちに少規模な品質の高い人手アノテーションデータで再度学習するように段階的にモデルの学習が行われたことがある．このような段階的な学習が F 値向上の要因であることが考えられる．

5.　他言語データを用いた学習

本章では複数の言語のデータセットを用いてモデルの学習を

(注4)：https://github.com/cl-tohoku/bert-japanese

表 3　日本語データを用いた学習の F 値 (疑似データ・HA1 で学習，HA2 でテスト)

		直喩	非直喩	平均
疑似	pseudo	0.604	0.701	0.653
人手	HA1	0.701	0.814	0.758
拡充	HA1 + 100	0.686	0.805	0.745
	HA1 + 10000	0.649	0.756	0.702
追加学習	pseudo → HA1	**0.732**	**0.831**	**0.782**

表 4　日本語データを用いた学習の F 値 (疑似データ・HA2 で学習，HA1 でテスト)

		直喩	非直喩	平均
疑似	pseudo	0.731	0.759	0.745
人手	HA2	0.729	0.811	0.770
拡充	HA2 + 100	0.416	0.798	0.757
	HA2 + 10000	0.670	0.731	0.697
追加学習	pseudo → HA2	**0.809**	**0.872**	**0.840**

行った結果を示す．5.1 節では本実験で用いるマルチリンガル学習の枠組みについて説明する．5.2 節で実験の設定について説明し，5.3 節で実験結果について示す．

5.1　マルチリンガル学習

本研究のマルチリンガル学習はマルチタスク学習の枠組みを用いる．マルチタスク学習は転移学習の一種とされ，関連する複数のタスクに関して有用な情報を共有しながら学習を行うことで，モデルの汎化性能を高め，単一モデルの学習よりも精度の向上が期待できることが知られている [10].

マルチタスク学習を行う方法は様々存在するが，本実験では各タスクにそれぞれ独立したモデルを持つソフトパラメータシェアを採用する．また，分類モデルとして BERT を用い，BERT が持つ 12 層の Transformer 層の一部のパラメータを共有する形でマルチタスク学習を行う．

本タスクにおけるマルチリンガル学習モデルの概要を図 2 に示す．マルチタスク学習では異なるタスクを解くモデルを図 2 のように並列に学習させるが，本タスクではそれぞれ異なった言語のモデルを並列に並べ，それぞれの言語で同様のタスクを解かせる．多言語で並列して学習することにより，元の言語に存在しない直喩の特徴をモデルが学習し精度が向上することを期待する．また，本タスクにおいては Pahari ら [15] の研究に倣い，BERT の 2〜8 層のみのパラメータを共有する．

5.2　実験設定

本実験では学習する言語に応じて 3 種の BERT モデルを用いる．日本語データに対しては前節と同様に *cl-tohoku/bert-japanese*[5]を用いる．英語データセットに対しては *bert-base-uncased*[6]を用いる．また，中国語データセットに対しては，*bert-base-chinese*[7]を用いる．

次に，本実験で用いる BERT のハイパーパラメータ設定を説明する．日本語 BERT に関しては 4.2 節と同様である．英語 BERT，中国語 BERT は共に，最適化関数に Adadelta，損失関数に BCELoss を用いる．学習率は英語 BERT は 5e-3，中国語 BERT は 1e-3 とし，重み減衰は共に 1e-5 に設定する．入力の最大トークン長は 64，バッチサイズは 64 で両モデル共通である．最大エポック数は 60 として Early Stopping を導入し，すべてのモデルにおいて 10 エポック間モデルの更新がなかった場合，学習を打ち切る．

（注5）：https://github.com/cl-tohoku/bert-japanese
（注6）：https://huggingface.co/bert-base-uncased
（注7）：https://huggingface.co/bert-base-chinese

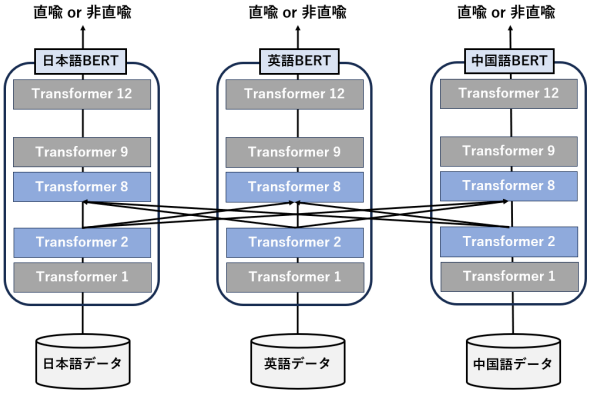

図 2　マルチリンガル学習の概略

本実験での英語データと中国語データの取り扱いについて述べる．*MSD* では英語，中国語共に 9 : 1 に分割し，それぞれの訓練データと開発データとする．CSR_{zh} では，公開されたデータセットがあらかじめ分割されているため，データ分割に関してはデフォルトの設定に倣う．なお，日本語データを用いる場合の設定は前節と同様である．また，訓練データに関しては前節と同様に，全てのデータでランダムなダウンサンプリングを行い，直喩と非直喩のデータ数を揃える．訓練データを選定してモデルの学習を行う過程を 3 回行い，平均値を算出する．

各モデルに関して 2 言語及び 3 言語の複数の組み合わせを検討し，比較する．日本語データに関しては前節で最も高い F 値を記録した pseudo → HA1，pseudo → HA2 を用いる．このとき，疑似データによる学習は単一のモデルで行われ，人手データによる学習の際にマルチリンガルでの学習が行われる．

5.3　実験結果

本論文では日本語データの分類結果のみを表 5 と表 6 に示す．テストデータには前節と同様に，人手アノテーションデータのうち，学習に用いられていないものを用いる．また，表の値はそれぞれ直喩，非直喩の F 値とその平均 F 値である．表中の太字はそれぞれの結果のうち，最高 F 値を示す．

HA1 を用いたマルチリンガル学習においては，2 言語間で学習したモデルにおいて F 値の向上が見られた．中国語のエッセイデータを基に構築された CSR_{zh} と相互学習したモデルが一番良い F 値を得た．Zhu ら [11] の研究では，表層情報の似ている日本語と中国語で相互学習を行うことによってより良い結果が得られることが示されており，今回の結果も同様の傾向が見られる．また，3 言語間で学習したモデルに関しては精度が低下する結果となった．これは，日本語モデルが 3 つのモデルの情報を処理できず，学習が上手くいかなかったと考えられる．

表 5　マルチリンガル学習の F 値 (HA1 で学習，HA2 でテスト)

		直喩	非直喩	平均
1 言語	p → HA1_A	0.732	0.831	0.782
2 言語	(p → HA1) + MSD_{en}	0.747	0.850	0.798
	(p → HA1) + MSD_{zh}	0.756	0.860	0.808
	(p → HA1) + CSR_{zh}	**0.763**	**0.866**	**0.814**
3 言語	(p → HA1) + MSD_{en} + MSD_{zh}	0.710	0.822	0.766
	(p → HA1) + MSD_{en} + CSR_{zh}	0.705	0.811	0.758

表 6　マルチリンガル学習の F 値 (HA2 で学習，HA1 でテスト)

		直喩	非直喩	平均
1 言語	p → HA2_A	**0.809**	0.872	**0.840**
2 言語	(p → HA2) + MSD_{en}	0.803	**0.877**	**0.840**
	(p → HA2) + MSD_{zh}	0.797	0.875	0.836
	(p → HA2) + CSR_{zh}	0.801	**0.877**	0.839
3 言語	(p → HA2) + MSD_{en} + MSD_{zh}	0.799	0.875	0.837
	(p → HA2) + MSD_{en} + CSR_{zh}	0.793	0.874	0.833

一方で，HA2 を用いたマルチリンガル学習においては，2 言語間の学習と 3 言語間の学習の全ての場合で日本語のみの学習よりも F 値の向上が見られなかった．この結果に関しては，HA2 の方が HA1 よりもデータサイズが大きく単体のモデルでも充分に学習が行われていたことが原因だと考えられる．充分に学習された日本語モデルはマルチリンガル学習によって追加で得られる情報の恩恵が乏しかったと推察される．

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別に取り組んだ．また，機械学習のためのアノテーションデータの不足を補うために疑似ラベリングデータを用いた学習，マルチリンガル学習の 2 つを取り入れ直喩文判別の更なる精度向上を目指した．

1 つ目の疑似ラベリングデータを用いた学習では，機械翻訳の日英翻訳を用いて変換した英文を基に文章に対して疑似ラベリングを行い，データセットを構築した．結果として，疑似ラベリングデータでファインチューニングを行った BERT に対して更に人手アノテーションデータでファインチューニングを行うことで最高 F 値を得た．このことから，大量の疑似データで学習した後に少量の品質の良い人手アノテーションデータで学習するような段階的な学習が有効であることが示された．

2 つ目のマルチリンガル学習では，直喩文判別を英語と中国語で解くモデルと相互学習を行った．結果として日本語での直喩文判別タスクでは，HA1 とエッセイ文を基とする中国語データである CSR_{ch} との相互学習で最高 F 値を得た．また，もう一つの日本語アノテーションデータである HA2 を用いた実験では F 値の向上が見られなかった．これは，日本語 BERT 単体で判別を行った場合の F 値が HA1 よりも高かったこと，HA1 よりもデータ数が多く学習が充分に行われていたことが原因として考えられる．

今後の課題として，疑似ラベリングデータの精査が挙げられる．本研究では疑似ラベリングの特徴語として英語の前置詞 “like” を使用したが，より適切な判断基準を追加することができれば疑似ラベリングの精度を上げることができる．実験に用いる疑似ラベリングデータの品質が向上すれば，更に直喩文判別の精度が向上することが期待できる．また，直喩文判別の後段タスクとして直喩文から喩辞と被喩辞を抽出するタスクが存在する．今後は前述のタスクにも取り組み，直喩文からの更なる情報抽出に努めたい．

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Jintaro Jimi and Kazutaka Shimada. Pseudo data acquisition using machine translation and simile identification. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 391–396, 2022.

[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] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 12, pp. 5586–5609, 2022.

[11] Chencheng Zhu, Niraj Pahari, and Kazutaka Shimada. Multilingual symptom prediction by simultaneous learning using bert. In *2023 International Conference on Asian Language Processing (IALP)*, pp. 100–105, 2023.

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

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

[14] Longxuan Ma, Weinan Zhang, Shuhan Zhou, Churui Sun, Changxin Ke, and Ting Liu. I run as fast as a rabbit, can you? a multilingual simile dialogue dataset. *arXiv preprint arXiv:2306.05672*, 2023.

[15] Niraj Pahari and Kazutaka Shimada. Multi-task learning using bert with soft parameter sharing between layers. In *12th SCIS and 23rd ISIS*, pp. 1–6. IEEE, 2022.