修　　士　　論　　文

知識ベースを活用した
汎用多言語モデル

指導教員　　越前功 教授

東京大学大学院情報理工学系研究科
電子情報学専攻

氏 名　48-206443 西川荘介

提 出 日　　令和四年 01 月 25 日

概 要

自然言語処理分野では近年、複数言語における自然言語処理モデルの学習・管理を容易にするため、単一のモデルで複数の言語の入力に対応する多言語モデルが盛んに研究されている。しかしこれらのモデルは高い性能を達成するために、特定のタスク用に作成された教師データや翻訳データなど、言語によっては十分に手に入らない貴重な言語資源に依存している。本稿ではこの問題を緩和するため、インターネット上に集積されることで幅広い言語で容易に手に入る、あらゆる物事（エンティティ）に関する情報が構造化された知識のデータベース（知識ベース）が、多言語モデルの学習資源として有効であることを示した 2 種類の研究結果についてまとめる。

1 つ目は、ある言語の学習データで学習した多言語モデルを他の言語に追加学習なしで転用するゼロショット言語間文書分類タスクに対して、知識ベースが有効であるか検証した研究である。この研究では文書データから文書に関連のあるエンティティを抽出し、これらを分類タスクにおける追加の入力特徴量として用いる手法を提案する。文書中に登場するエンティティは単語とは異なり曖昧性がなく文書のトピックを捉えることができ、さらに言語に依存せず定義されるという性質があるため、これらを特徴量として用いることで言語間文書分類の性能が改善されることが期待される。実験では、複数の言語間文書分類タスクにおいて、提案手法が最高性能を達成した。

2 つ目は多言語モデルにおいて、複数言語の文の意味を数値ベクトルで表す多言語文表現学習に対して知識ベースが有効であるか検証した研究である。この研究では、Wikipedia における各文からその文中に登場するハイパーリンクに対応するエンティティを推定するタスクにより、多言語文表現を学習する手法を提案する。この手法では様々な言語の文が言語に非依存なエンティティを軸に共通のベクトル空間上に埋め込まれるため言語に非依存な文表現が学習されることが期待される。実験では多言語文表現を評価するいくつかの自然言語処理タスクにおいて提案手法が既存の教師なし多言語文表現モデルと比較し優れていることが確認された。

以上の 2 つの研究では知識ベースを多言語モデルの学習資源として利用する有効性を示した。知識ベースは現在も拡張が進められており将来的にはより幅広い言語で豊富に手に入る重要な言語資源となることが考えられ、このような資源の活用に焦点を当てることは極めて重要である。また、本研究の成果は今後知識ベースを含めた様々な言語資源を活用したより良い多言語モデル構築への足掛かりとなる。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

自然言語をコンピュータで処理する自然言語処理 (Natural Language Processing; NLP) の技術は、近年の計算機の演算能力の向上やインターネットの普及によるビッグデータの形成に伴う深層学習技術の発展により、著しい発展を遂げている。自然言語処理技術は検索サイトや翻訳サイトを始めとする我々が日常的に利用するサービスにも応用されており、現代社会に欠かせない重要な技術の一つである。

何らかのタスクを解く自然言語処理モデルを学習する場合、多くの場合は人手によってラベルが付与されたアノテーションデータを学習データとして用いる。しかし、複数の異なる言語話者の利用を想定した自然言語処理モデルを構築する場合、2022 年 1 月現在世界で約 7,000 の言語が存在することを考慮すると、個別にアノテーションデータの構築やモデルの学習・管理を行うことは膨大なコストがかかるため非現実的である。このような背景から 1 つの自然言語処理モデルで複数の異なる言語を扱うことが可能な多言語自然言語処理モデル（多言語モデル）が提案されている。深層学習技術によるモデルの学習はモデルの内部に密な連続値ベクトルを隠れ表現として持ち、これらの行列演算によって実現される。このため、多言語モデルの構築にあたっては入力系列を言語に依存しない入力ベクトルに変換さえできればあとは単言語の場合と同様に end-to-end での単一のモデルの学習・推論が容易に可能である。従来煩雑な処理が必要だった多言語モデルは、この深層学習技術との親和性の高さにより容易に構築することが可能となり近年盛んに研究されている。

近年の深層学習技術を用いた多言語モデルの構築では、まず大規模な複数の異なる言語のテキストコーパスを用いて言語を区別せず教師なし学習を行うことで、事前に多言語モデルを学習する。このように構築された事前学習済み多言語モデルに対して、特定のタスク用の教師データや対訳コーパスなどのアノテーションデータを用いて追加の学習を行うことで、特定タスクに対応する多言語モデルが構築される。しかし上述のような何らかのタスクや言語に依存する言語資源は言語によっては量に限りがあるため、多言語モデルの性能を引き出せない場合がある。本稿では、高性能な多言語モデルを学習する言語資源として Wikipedia などの知識ベース資源が有効であるかを検証した 2 つの研究についてまとめる。

1.2　本研究の貢献

本研究では、多言語モデルにおけるゼロショット言語間文書分類タスクと多言語文表現学習において、知識ベースを活用する有効性を検証する。

1 つ目のゼロショット言語間文書分類における研究では、分類対象の文書に登場するエンティティを抽出し、これらからエンティティの該当文書への関連度に応じて重みを算出する注意機構に基づくエンティティベースの特徴量を計算し、これを分類タスクの入力特徴量として利用する手法を提案する。この手法では該当文書から得られる単語ベースの特徴量に加えて、（1）言語に非依存であり（2）文書のトピックをよく捉えるエンティティの特徴量を利用するため、言語間文書分類タスクの性能が向上することが期待される。実験では 3 つのゼロショット言語間文書分類タスクにて、提案手法がベースラインモデルを上回る性能を達成した。

2 つ目の多言語文表現学習における研究では、ベクトル空間上で文をその文に関連するエンティティに近づけ、関連しないエンティティとは遠ざけることで文表現を学習する手法を提案する。エンティティは言語に依存せず定義されるため、この手法によって文表現が言語非依存になることが期待される。また、文を関連するエンティティに近づけるため、似た意味を持つ文が、類似した文表現になることが期待される。実験では（1）複数言語の学習データで文表現を学習する多言語の設定、（2）英語の学習データで文表現を学習する単言語の設定で提案手法により文表現を学習し、それぞれの設定において様々な自然言語処理タスクで評価を行った。その結果、提案手法により学習した文表現は多くのタスクでベースラインモデルを上回り、言語に非依存で文の意味を上手く捉えていることが確認された。さらに本研究では多言語文表現評価用のデータセットとして 16 言語、13 カテゴリからなる多言語文書クラスタリングデータセットを新たに構築した。

1.3 本研究の構成

本稿の構成について説明する。まず第 2 章では本研究の前提となっている汎用多言語モデルや言語間転移学習などの背景知識について説明する。次に第 3 章では文書から抽出した知識ベースエンティティを追加の特徴量として用いることで汎用多言語モデルにおけるゼロショット言語間文書分類の性能の向上を試みた研究について記述する。第 4 章では文とそれに関連するエンティティを用いた Contrastive learning を行うことで高性能な多言語文表現の学習を目指した研究について記述する。最後に第 5 章では本稿のまとめを行う。

第2章　背景知識

2.1　汎用言語モデル

汎用言語モデルとは事前に大量の文書データで学習された自然言語処理モデルであり、用途に応じた再学習を行うことで、該当タスクに対して高い性能を発揮することが可能なモデルである。本節ではまず、近年の汎用言語モデル構築のベースとなるモデル Transformer について解説し、その後、本研究に関わるいくつかの汎用言語モデルについて解説する。

2.1.1　Transformer

従来ニューラルネットワーク機械翻訳を代表とする系列変換タスクにおいては主に Recurrent Neural Network (RNN) や Long Short-term Memory (LSTM) [1]、Convolutional Neural Networks (CNN) を中心としたモデルが利用されてきた。近年、これらのモデルを利用せず後述する注意機構 (attention) を中心に構成された Transformer [2] が提案された。このモデルは RNN や LSTM のような再帰的なモデルのように逐次的にトークンを処理する必要がないため並列化が容易であり、学習が高速である。また、近年では自然言語処理における多くの分野でこの機構に基づいた手法が提案され、目を見張る成果が報告されている。以下ではこのモデルの詳細を説明する。

図 2.1, 2.2 に Transformer モデルの全体図を示した。

Transformer における注意機構は式 (2.1) 式によって表され、Scaled Dot-Product Attention と呼ばれる。

$$Attention(Q, K, V) = softmax(\frac{QK^t}{\sqrt{d_k}})V \tag{2.1}$$

ここで n は入力文の系列長を表す。また、$Q \in R^{n \times d_k}$、$K \in R^{n \times d_k}$、$V \in R^{n \times d_v}$ はそれぞれ query、key、value と呼ばれるベクトルであり、入力ベクトルに各層の重み行列 W_Q, W_K, W_V をかけることで生成される。この要素を用いて (2.1) 式に従い系列内でのトークン同士の関係性を表すような注意機構スコアが各トークンごとに計算される。また QK^t の各要素が大きくなりすぎて，逆伝播の softmax の勾配が極端に小さくなることを防ぐために、Q、K の隠れ層の次元である d_k の平方根の絶対値で除算している。

次に図 2.1, 2.2 中の Multi-Head Attention 層について説明する。この層は上述の Scaled Dot-Product Attention を１つのヘッドと見做しパラメータの異なる複数ヘッドを並列化したものであり、(2.2) 式、(2.3) 式で表される。

$$MultiHead(Q, K, V) = Concat(head_1 .....head_h)W^O \tag{2.2}$$

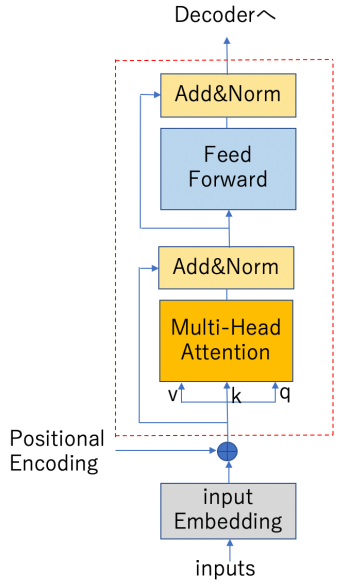


図 2.1: Transformer (エンコーダ)

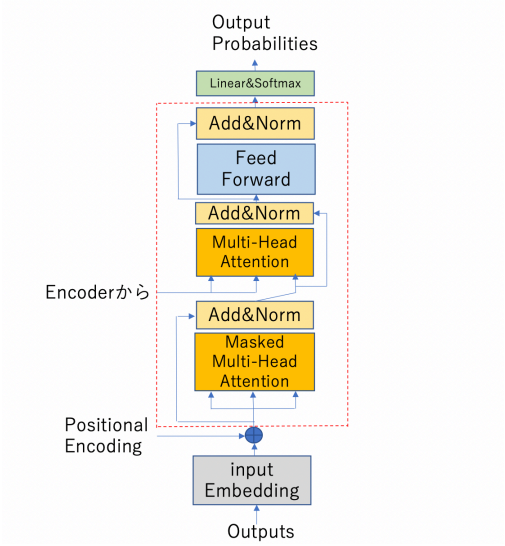


図 2.2: Transformer (デコーダ)

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

ここで $W_i^Q \in R^{d_{model} \times d_k}$、$W_i^K \in R^{d_{model} \times d_k}$、$W_i^V \in R^{d_{model} \times d_v}$、$W^O \in R^{hd_v \times d_{model}}$ はそれぞれ重み行列を表し、$Concat(\cdot)$ は連結を表す。この層では複数の異なるヘッド上で複数の潜在表現を処理するため、より豊かな情報が獲得できる。

Multi-Head Attention 層の結果は Position-Wise Feed-Forward Networks (Feed Forward) 層に入力される。この層は (2.4) 式で表され各トークンごとに処理される。

$$FFN(\boldsymbol{x}) = \max(0, \boldsymbol{x}W_1 + \boldsymbol{b}_1)W_2 + \boldsymbol{b}_2 \tag{2.4}$$

ここで $W_1 \in R^{d_{ff} \times d_{model}}$、$W_2 \in R^{d_{model} \times d_{ff}}$ は重み行列であり、$b_1 \in R^{d_{ff}}$、$b_2 \in R^{d_{model}}$ はバイアスを表す。また $\max(0, f(x))$ は活性化関数 ReLU を表す。

Transformer では図 2.1, 2.2 中にあるように Add&Norm 層と呼ばれる層が Multi-Head Attention 層もしくは Feed Forward 層の後に導入されている。Add&Norm 層の出力行ベクトル $\boldsymbol{y}$ は以下の式で表される。

$$\boldsymbol{y} = LayerNorm(\boldsymbol{x} + Sublayer(\boldsymbol{x})) \tag{2.5}$$

ここで $\boldsymbol{x}$ は入力行ベクトル、$Sublayer$ は Multi-Head Attention 層または Feed Forward 層の出力を表し、$LayerNorm$ は層ごとに正規化する Layer Normalization [3] を表している。入力の加算により入出力の差分を学習させること（Residual Connection [4]）で勾配消失を防ぎ、Layer Normalization により学習速度を向上させている。

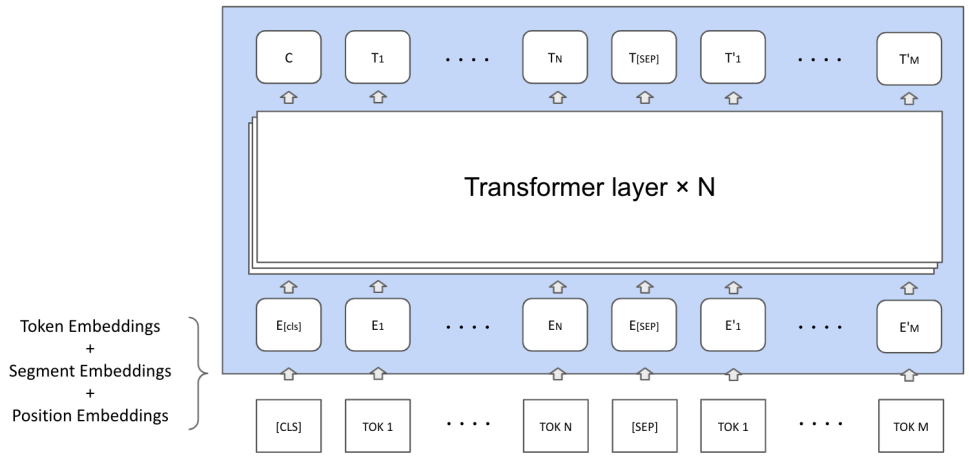


図 2.3: BERT の概要図

Transformer では RNN のような再帰的な構造を使用しないため、トークン間の前後関係の情報を加える必要がある。これを実現する手法として Positional Encoding が用いられ、入力の埋め込み行列に以下の式で表される値を要素ごとに加算することで位置情報を導入する。

$$PE_{(pos,2i)} = \sin \frac{pos}{10000^{\frac{2i}{d_{model}}}} \tag{2.6}$$

$$PE_{(pos,2i+1)} = \cos \frac{pos}{10000^{\frac{2i}{d_{model}}}} \tag{2.7}$$

ここで PE は Positional Encoding、pos はトークンの系列中での位置を表し、$2i$、$2i+1$ はベクトル内の次元の番号を表す。

上述の機構を用いて図 2.1、2.2 の Transformer モデルが構成される。点線で囲まれたブロックを N 層[1]重ねる。

また、Transformer では学習を安定させるため、最初に小さな学習率を設定し、そこからあるステップ数まで徐々に学習率を上げていき、その後に再度学習率を下げるウォームアップという手法が採用されている。Transformer では学習率が最初から大きいと Layer Normalization のパラメータが学習初期には安定せず学習が収束しない場合があり、これに対応するためにウォームアップが採用されている。

2.1.2 Bidirectional Encoder Representations from Trasnformers (BERT)

近年 GPU や TPU などの計算資源の発展により、大規模なデータセットで学習された数億規模のパラメータを持つモデルが続々と考案され、様々な自然言語処理タスクにおいて高い性能を発揮している。代表的なモデルとしては、LSTM モデルに基づく分散表現 Embeddings from Lanuage Models

[1]元論文 [2] では N=6 を採用している。

(ELMo) [5] や Transformer に基づく言語モデル Generative Pre-trained Transformer (GPT) [6, 7] などがある。しかし、これらの言語モデルの目的関数は一方向のトークン列の関係性の情報しか考慮されないため、あるトークンに関して前後の情報が考慮されず、文全体の情報が重要となる文レベルのタスクや Question Answering (QA) などのトークンレベルのタスクには向いていない。上述の問題を解決するため、Transformer に基づく大規模事前学習モデル Bidirectional Encoder Representations from Trasnformers (BERT) [8] が提案された。

BERT では Masked Language Model (Masked LM) と Next Sentence Prediction (NSP) と呼ばれる二つの手法によって学習される。Masked LM では系列中のトークンをランダムにマスクトークンと呼ばれる特殊トークン $[MASK]$ に置き換え、その部分を予測するタスクを解くことで学習する。実際に後述するファインチューニングを行う際はマスクトークンが登場しないため、事前学習時とファインチューニング時のギャップを和らげるため実際には以下のような処理によりマスクトークンを導入する。

- 15%の確率でトークンを選択する。それらのうち
  - 80%を $[MASK]$ に置き換える。
  - 残りの 10%をランダムなトークンに置き換える。
  - 残りの 10%を元のトークンのままにする。

さらに後述するファインチューニングで用いられる $[CLS]$ トークンをトークン系列の先頭に挿入する。また文の末尾または文の境界には $[SEP]$ トークンを挿入する。Masked LM によりあるトークンに関して双方向のトークンの情報を考慮された表現の学習が可能となる。

BERT ではさらに NSP という学習手法も導入している。これは二つの文が連続しているかどうかを判定するタスクである。実際には $[SEP]$ トークンで連結された二つの文を入力し、連続しているかどうかの二値分類タスクを学習する。NSP により QA や Natural Language Inference (NLI) などの二つの文の関係性を推定するタスクへの対応が可能となる。

BERT のアーキテクチャは Transformer のエンコーダを複数層重ねて構成される。このアーキテクチャにて約 330 億語を含む大規模コーパスを用いて上述の手法により事前学習が行われる。

BERT を個別のタスクに適用させる際、ELMo のように隠れ層の表現のみを特徴量として用いるのではなく、タスク用のモジュールをモデルに追加し、モデル全体のパラメータを更新する。このような学習手法はファインチューニングと呼ばれる。例えば文書分類タスクを代表とする文レベルのタスクを解く際は、$[CLS]$ トークンに対応する隠れ層の最終層の表現をタスク用モジュールへの入力とする。

BERT は複数の言語処理タスクを含む GLUE [9] や、代表的な QA のデータセット SQuAD [10] にて、論文公開当時の state of the art の性能を達成している。

2.1.3 RoBERTa

上述の BERT の発表後、BERT を改善した派生モデルがいくつか報告されている。Liu らの発表した RoBERTa [11] はその派生モデルの一つであり、BERT について分析を進め、以下に挙げるシンプルな改善で多くの自然言語処理タスクにおける性能改善に成功した。

1. より大規模なデータセットかつ大きいバッチサイズにてより長時間の学習を行う
2. NSP を使用しない
3. より長い文章を用いて学習を行う
4. 入力するごとに動的に入力トークン系列のマスクパターンを変化させる

Liu ら [11] は実験にて上記の改善により BERT の性能が向上することを検証している。3 については、BERT では Masked LM の際、$[MASK]$ トークンの位置を事前学習前に一度ランダムに決めるが、それでは異なるエポックで同じマスクパータンが登場してしまう。この点を改良するために入力するごとに動的に入力トークン系列のマスクパターンを変化させている。その結果 GLUE や、SQuAD を含めたいくつかのデータセットにて BERT を超える性能を達成した。

2.2 言語間転移学習

2.2.1 ゼロショット言語間転移学習

深層学習技術の台頭により、自然言語処理の各分野のタスクにおいて英語などの資源が豊富な言語を中心に目覚ましい性能の向上が報告されている。しかし、実世界に存在する約 7000 の全ての言語において人間によるアノテーションがなされた学習データを用意し、機械学習モデルを構築するのは非現実的である。その結果、特定の言語群に特化した技術が主に発展し、自然言語処理技術の言語格差が進んでいる。

この問題を解決するために、ゼロショット言語間転移学習と呼ばれる、資源の豊富な言語（原言語）におけるラベル付きデータで学習したモデルを、（特に言語資源が乏しい）他の言語（目的言語）におけるテストデータの推論に利用する手法が盛んに研究されている。ここでゼロショットと呼ばれるのは、目的言語のラベル付きデータを一切使わずに目的言語におけるテストデータの推論を行うためである。

ゼロショット言語間転移学習を実現するにあたってこれまでいくつかの手法が提案されてきている。Wan らは原言語のラベル付きデータや目的言語のラベルなしデータに対して、機械翻訳で翻訳することで構築した学習データセットから感情分析の分類器を学習した [12]。また、機械翻訳の代わりに翻訳文の集合データである対訳コーパスから言語間転移学習を実現する手法も提案されている。代表的な手法としては Meng らの cross-lingual mixture model (CLMN) [13] がある。このモデルでは対訳コーパスを活用し、原言語文から目的言語文を復元するような学習を行うことで対訳コーパス内に存在するがラベル付きデータには存在しない語彙を推測し言語間の差異を解消する。

しかし、上述の手法の学習に必要となる大規模な対訳コーパスは全ての言語対で容易に手に入るわけではない。すなわち、大規模な対訳コーパスが手に入るような資源が豊富な言語に対してしか適用できない。これはゼロショット言語間転移学習の目的の一つである低資源言語における自然言語処理モデルの活用の促進を実現しているとは言い難い。

その他の手法として多言語埋め込み表現を用いたゼロショット言語間転移学習の手法がいくつか提案されている。異なる言語においてそれぞれ独立に Skip-gram [14] などにより学習された単

語埋め込み表現上において、似た意味を表す単語の埋め込み表現の類似度が高くなるとは限らない。そこで言語に依存せず、似た意味を表す単語埋め込み表現の類似度が高くなるように、複数言語の単語埋め込み表現を何らかの方法によって共通空間に写像した表現を多言語埋め込み表現 (Multilingual word embedding) と呼ぶ。多言語埋め込み表現の代表的な学習手法として、Mikolov ら [15] の線形写像によるものがあり、単語レベルの翻訳辞書を用いて (2.8) 式のように平均二乗誤差を最小化することである原言語の単語埋め込み表現空間から他の目的言語の単語埋め込み表現空間への変換行列 W を学習する。

$$\omega_{MSE} = \sum_{i=1}^{n} \|Wx_i^s - x_i^t\|^2 \tag{2.8}$$

ここで x_i^s, x_i^t は単語書にある単語ペアの i 番目のそれぞれの言語の単語埋め込み表現を表す。また多言語埋め込み表現の学習手法には対訳コーパスで学習する手法 [16] や、単語の類似度分布を利用することによって初期の単語レベル翻訳辞書を作成し、その辞書を利用して再度式 (2.8) 式に従って変換行列を学習することを繰り返す教師なしの手法 [17] も提案されている。

多言語埋め込み表現を用いた言語間転移学習ではモデルに対し、入力文の単語群を多言語埋め込み表現に変換する層を最初に導入する手法が一般的である [18]。この手法において学習時は原言語の学習データを入力データとし、多言語埋め込み表現への変換層のパラメータのみ固定しながら該当タスクを解くモデルを学習する。次に推論時は従来通り目的言語におけるテストデータを入力し、推論結果を得る。この手法ではモデルへの入力文の言語が学習時と推論時で異なっていても、多言語埋め込み表現上で言語の違いが吸収され目的言語のタスクを解けることが期待される。

しかし、多言語埋め込み表現を用いた手法においても依然として人手によって作成された対訳コーパスや単語レベルの翻訳辞書に依存しているものも多く、そのような資源が手に入る言語にしか適用できないという問題は残る。また対訳コーパスや翻訳辞書を利用しない多言語埋め込み表現学習手法も提案されているが、多義語の存在や文法の違いにより言語対で全ての単語が一対一対応しているわけではないため, 単語レベルで対応関係をとるのは限界があり、言語間転移の性能が実用的ではない場合が多い。

2.2.2 多言語汎用言語モデル

近年では言語間転移学習手法の多くは事前に多言語の大規模コーパスで学習された事前学習済み汎用多言語モデルを用いて実現され、様々なタスクで高い性能を発揮している。

代表的なモデルの一つである Multilingual BERT (mBERT) [8] はその名の通り、BERT を多言語に拡張したモデルである。mBERT は Wikipedia にてページ数が特に多い 104 の複数言語における Wikipedia コーパスを学習データとして用い、BERT と同様の手法で Transformer のエンコーダを複数重ねたモデルを学習することで構築される。

mBERT の興味深い点として学習データの言語と推論時の言語で語彙の重複がない場合でも高い性能での言語間転移が可能である点が挙げられる。Pires ら [19] の実験では POS tagging と呼ばれる品詞を推定するタスクにおいて mBERT をウルドゥー語の学習データでファインチューンを行うと、ウルドゥー語と全く語彙の重複していないヒンディー語の推論データにおいて、ヒンディー

語の POS タグ教師データを一切モデルは学習していないのにも関わらず 91%の正解率を達成することを確認している。さらに Pires らは mBERT において英語と日本語間など類型学的類似性が低い言語間においては言語間転移タスクの性能が低下することを示しており、mBERT は言語間の語彙の違いは吸収できるが、語順の違いは上手く吸収できていないと主張している。

その他のモデルとして RoBERTa を多言語に拡張したモデルである XLM-RoBERTa (XLM-R) [20] がある。XLM-R は mBERT の学習時に利用した Wikipedia コーパスよりもさらに大きい約 2.5TB の 100 言語で構成される CommonCrawl コーパス[2]を用いて Masked LM により学習され、Cross-lingual Question Answering [21] や Cross-lingual Natural Language Inference (XNLI) [22] などを含む様々な言語間転移タスクにおいて mBERT を超える性能を発揮した。また、XLM-R を提案した Conneau ら [20] は事前学習時の言語数を一定数以上増加させると多言語モデルの容量に限界がきて、個々の言語におけるタスクや言語間タスクの性能が低下してしまうことなど多言語モデルに関する重要な知見をいくつか示している。

これらの事前学習済み汎用多言語モデルでは対象トークンにおいて固有のベクトルを学習するのではなく、前後のトークン情報を考慮した文脈付き分散表現を学習することができる。この文全体の情報が埋め込まれた分散表現を利用することで、多義語の存在や文法の違いによる言語間の差異がある程度緩和されることが期待される。また BERT や RoBERTa における Masked LM や NSP などの事前学習手法は教師なしで行われる学習であるため、アノテーションされたデータを利用する必要がなく、インターネット上で公開されているテキストコーパスを大量に用いた学習が可能となる。従ってアノテーションデータが乏しい低資源言語への言語間転移により適した学習手法と言える。

2.3 知識ベースの自然言語処理への活用

2.3.1 Wikipedia2Vec

知識ベースとは組織化されインターネット上に集積された知識のデータベースを表す。代表的な知識ベースとしては Wikipedia[3]や Wikidata[4]などがあり、日々有志の手によって拡張・管理されている。特に Wikipedia は膨大であり、最も多い英語版 Wikipedia では 2022 年 1 月時点で 5 億件以上のページが存在している。

人間は知識ベースに存在するような実世界の膨大な知識を前提として行動を決定する。機械学習モデルは単なるテキスト情報からのみでは、そのような実世界の知識は学習できず、人間のような柔軟な判断ができない場合がある。そのような背景から知識ベースにおける情報を自然言語処理に取り込む研究が盛んに行われている。知識ベースは、人物や何らかの作品、出来事などあらゆる物事（エンティティ）を説明するページにより構成される (図 2.4)。また、エンティティに関する情報が集約された低次元ベクトルはエンティティ表現と呼ばれる。近年のエンティティを活用した自然言語処理の多くの研究では、このエンティティ表現を何らかの形で自然言語処理モデルに導入する。

[2] https://commoncrawl.org/about/
[3] https://en.wikipedia.org/wiki/Main_Page
[4] https://www.wikidata.org/wiki/Wikidata:Main_Page

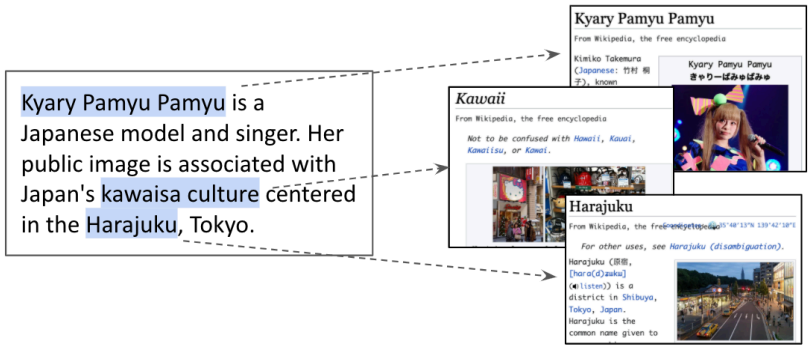


図 2.4: 文書中に現れるエンティティの例。

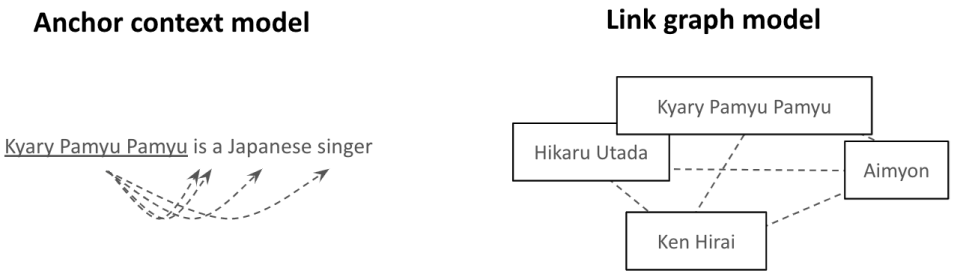


図 2.5: Wikipedia2Vec の学習モデル

2.3.2 Wikipedia2Vec

エンティティ表現を学習する代表的な手法として Wikipedia2Vec [23] がある。Wikipedia2Vec は Wikipedia のハイパーリンクで接続されたエンティティデータを利用し、以下の三つの目的関数を最適化することで学習される。また、これらの目的関数のイメージ図を図 2.5 に示す。

$$P(o_c|o_i) = \frac{\exp({\mathbf{V}_{o_i}}^\top \mathbf{U}_{o_c})}{\sum_{o \in O} \exp({\mathbf{V}_{o_i}}^\top \mathbf{U}_o)}, \tag{2.9}$$

$$\mathcal{L}_w = -\sum_{i=1}^{N} \sum_{-c \le j \le c, j \neq 0} \log P(w_{i+j}|w_i), \tag{2.10}$$

$$\mathcal{L}_a = -\sum_{(e_i, Q) \in A} \sum_{w_c \in Q} \log P(w_c|e_i), \tag{2.11}$$

$$\mathcal{L}_e = -\sum_{e_i \in E} \sum_{e_o \in C_{e_i}} \log P(e_o|e_i), \tag{2.12}$$

ここで (2.9) 式は Skip-gram モデル [14] の条件付き確率を表し、o はアイテム（エンティティまたは単語）を表し、o_c は o の周辺アイテムを表す。また、$\mathbf{V}_o \in \mathbb{R}^d$、$\mathbf{U}_o \in \mathbb{R}^d$ はそれぞれ o に対応

する embedding を表す。(2.10) 式は Skip-gram と同様に単語が与えられた際、ある単語の周辺の単語を推測する目的関数を表し、単語埋め込み表現を学習する。(2.11) 式は Anchor context model と呼ばれ、ハイパーリンク先のエンティティからその周辺の単語を推測することで似た意味を持つエンティティと単語の埋め込み表現をベクトル空間で近づける。(2.12) 式は Link graph model と呼ばれ、ハイパーリンクで接続されたエンティティのリンクグラフにおいて、近傍のエンティティを推測することでエンティティ同士の関係性を学習する。

Wikipedia2Vec によるエンティティ表現学習ライブラリはオープンソースとして公開されており[5]、容易に利用可能である。Wikipedia2Vec はエンティティリンキング [24] や固有表現抽出 [25]、テキスト分類 [26] など、様々な自然言語処理タスクで利用されている。

[5]https://github.com/wikipedia2vec/wikipedia2vec

第3章　ゼロショット言語間文書分類のための多言語エンティティモデルの構築

3.1　序論

近年、資源の豊富な言語の教師データで学習した自然言語処理モデルを、他の特に資源が乏しい言語に言語の教師データを一切使わずに転用する、ゼロショット言語間転移が盛んに研究されている。ここで教師データを利用した言語を原言語、転用先の言語を目的言語と呼ぶ。近年では大規模なコーパスを用いて事前に教師なし学習を行う汎用多言語モデルの登場により、ゼロショット言語間転移学習は大きな発展を遂げた。

しかしながら近年のいくつかの研究では、汎用多言語モデルによるゼロショット言語間転移は必ずしも上手くいかないことが報告されている [27, 28]。例えば Lauscher ら [28] の研究では、言語間転移の様々なタスクの性能と（1）言語間の類似度[6]、（2）事前学習時のデータサイズとのそれぞれの相関を計測することで、原言語と類似していない言語や事前学習時のテキスト量が少ない言語への言語間転移の性能が低下することを明らかにした。これらの研究は、該当文書の単語情報のみを利用したゼロショット言語間転移の限界を示唆している。

本研究では、言語に非依存な知識ベースに紐づくエンティティを汎用多言語モデルに組み込むことでゼロショット言語間文書分類の性能を向上させる、多言語エンティティモデル (a multilingual bag-of-entities model, M-BoE) を提案する。知識ベースのエンティティは単語（テキスト情報）とは異なり文書の曖昧性のないセマンティクスを捉えることができ、いくつかの 文書分類タスクに活用されてきた [30, 31, 32, 33, 26]。我々のモデルは特に Wikidata 知識ベースのエンティティを入力の特徴量として用いることで汎用多言語モデルを拡張する（図 3.1）。Wikidata 知識ベースは、同一の概念を表すエンティティ (e.g., *Apple Inc.,* Е п ъ л *,* アップル) に対して言語に依存しないユニークな識別子 (e.g., Q312) が割り当てられているという特徴がある。

多言語エンティティモデルではまず、文書データから抽出された Wikipedia エンティティをそれぞれ対応する Wikidata エンティティに変換し、それらのエンティティ表現の重みつき和を計算することでエンティティベースの文書表現を得る。この重みは、山田らや Peters らの研究 [26, 34] に基づき、該当文書に関連するエンティティ表現を優先的に考慮するような注意機構により獲得される。次に、エンティティベースの文書表現と汎用多言語モデルから得られるテキストベースの文書表現を足し合わせ、分類タスクを解く線形分類器に入力する。Wikidata エンティティの語彙や表現は言語を渡り共有されているため、原言語で学習したエンティティの特徴量を直接複数の目的言語へ転移させることが可能となる。

[6] Lauscher ら [28] は様々な言語の特徴をベクトルにエンコードする LANG2VEC [29] を用いて言語間の類似度を計算している

実験では、汎用多言語モデルのベースモデルとして多言語 BERT [8] と XLM-R [20] を用い、このモデルを拡張することで多言語エンティティモデルを構築した。また 3 つの多言語文書分類タスク（トピック分類タスクである MLDoc [35] と TED-CLDC [36]、エンティティ型推定タスクである SHINRA2020-ML [37]）にて多言語エンティティモデルを評価した。その結果、多言語エンティティモデルは 3 つのタスクの全ての言語にてベースモデルを上回る性能を達成し、さらに MLDoc データセットでは既存の state of the art モデルより高い性能を発揮した。

3.2 関連研究

3.2.1 ゼロショット言語間文書分類タスク

文書分類タスクとは文書や記事に対してラベルを割り振るタスクである。特に 1 つの文書に対して 1 つのラベルを割り振るタスクはシングルラベル分類と呼ばれ、1 つの文書に対して複数のラベルを割り振るタスクはマルチラベル分類と呼ばれる。また、ゼロショット言語間文書分類タスクとはある言語（原言語）のラベル付きの学習データを活用して、他の言語（目的言語）の文書分類を目的言語のラベル付き学習データを一切使わずに解くタスクである。2.2.1 節で述べたようにこのタスクは機械翻訳や対訳コーパスを活用する手法 [12, 13]、多言語埋め込み表現 [18] などによって解かれてきたが、近年では mBERT に代表される多言語汎用言語モデルによって解く手法が盛んに研究されており、目覚ましい成果が報告されている。

3.2.2 追加データを利用したゼロショット言語間転移学習の手法

いくつかの既存研究では目的言語の追加データを利用して言語間転移学習の改善に取り組んでいる。Lai ら [38] は目的言語のラベルなしコーパスを利用して Masked LM や Unsupervised Data Augmentation [39] などの手法を用いて目的言語とのドメインのギャップを埋めるような学習手法を提案している。Keung ら [40] は目的言語のラベルなしコーパスを用いて、目的言語と原言語の隠れ表現から言語が特定できなくなるように敵対的な学習を行うことで mBERT の言語間転移学習の性能を向上させている。Dong ら [41, 42] や Eisenschlos ら [43] は目的言語のラベルなしコーパスに対して、原言語で一度学習したモデルからラベルを推定（pseudo-labeling）することでデータ拡張を行っている。Conneau ら [44] は追加の対訳コーパスを用いて、対訳コーパスを連結した文を入力とし、Masked LM を学習する Translation Language Modeling を提案している。

しかし、これらの学習手法は目的言語ごとの追加のコーパスを必要としている。さらにそれらの言語資源で学習した言語にしか多言語モデルを適用することができない。これらの学習手法とは異なり、多言語エンティティモデルでは追加のコーパスを必要とせず、一つの汎用多言語モデルの学習で複数の目的言語における文書分類の性能を向上させることができる。さらに多言語エンティティモデルは汎用言語モデルの内部構造を改変することなく、外部からエンティティ表現を導入するような単純な実装なため、既存の汎用言語モデルや上述した学習手法に対して容易に適用することができる。

3.2.3 エンティティによる汎用言語モデルの改善

いくつかの既存手法では事前学習段階でエンティティを汎用言語モデルに組み込む手法を提案している。ERNIE [45] や KnowBert [34] では事前学習済みエンティティ表現を挿入する形で汎用言語モデルに知識情報を組み込む。LUKE [46] や EaE [47] では事前学習時にエンティティ表現も同時に学習する。しかし、これらのモデルでは単言語の自然言語処理タスクの性能向上を目的としており、巨大なコーパスによる長時間の事前学習が必要となる。提案手法の多言語エンティティモデルでは事前学習を行うことなく、エンティティ情報を fine-tuning 時に動的に汎用多言語モデルに組み込む。

また、単言語の汎用言語モデルの自然言語処理タスクの性能を向上させるため、事前学習の後にエンティティ情報を組み込む手法を提案している手法も存在する。Ostendorff ら [48] は汎用言語モデルから得られる文脈付き表現と著者エンティティを表す知識グラフ表現を連結させたものを本分類タスクの特徴量として活用している。E-BERT [49] では知識ベースエンティティを文中のエンティティ名の隣に挿入することでエンティティ関連のタスクでの高い性能を達成している。その際、エンティティへはエンティティ表現を割り振るが、エンティティ表現空間から BERT のトークンベクトル空間へ写像を行うことでエンティティ空間と BERT のトークンベクトル空間の差異を緩和している。Verlinden ら [50] はスパン表現と知識ベースのエンティティ表現を組み合わせた表現を BiLSTM ベースの情報抽出モデルの内部に組み込んでいる。

これらのモデルとは異なり、多言語エンティティモデルは言語に非依存なエンティティ表現と汎用多言語モデルを組み合わせることでゼロショット言語間文書分類タスクの性能の向上を目指す。

3.2.4 エンティティを用いた文書分類の手法

いくつかの既存研究ではエンティティを文書分類タスクに活用している。Explicit semantic analysis (ESA) はその代表的な手法であり、各要素がある文書の各エンティティへの関連度スコアであるスパースなベクトルを用いてエンティティベースの文書表現を作成している [30, 31, 32]。さらに Song らは ESA で言語間文書分類タスクを扱うために ESA を拡張した cross-lingual explicit semantic analysis (CLESA) [33] を提案した。CLESA では Wikipedia の言語間リンクを用いて、原言語と目的言語において共通する Wikipedia エンティティを用いて ESA と同様にスパースなベクトルを計算する。

提案手法は CLESA とは異なり、state of the art である汎用多言語モデルを言語に非依存な Wikidata エンティティで計算されるエンティティベースの文書表現で拡張することで言語間文書分類タスクの性能向上を目指す。

提案手法に最も関連性の高い手法として山田らの提案した neural attentive bag-of-entites (NABoE) モデル [26] がある。NABoE モデルは文書中から Wikipedia エンティティを抽出し、さらに抽出されたエンティティの中で文書に関連する重要なエンティティを優先する注意機構によりエンティティ表現の重みつき平均を算出し、これを用いて単言語での文書分類タスクを解いている。多言語エンティティモデルはこの NABoE モデルの拡張モデルであり、（1）Wikidata 知識ベースを用いることで言語に依存しないエンティティ表現を活用している点と（2）エンティティベースの文書表現を大規模なコーパスで学習された汎用多言語モデルと組み合わせている点で異なる。

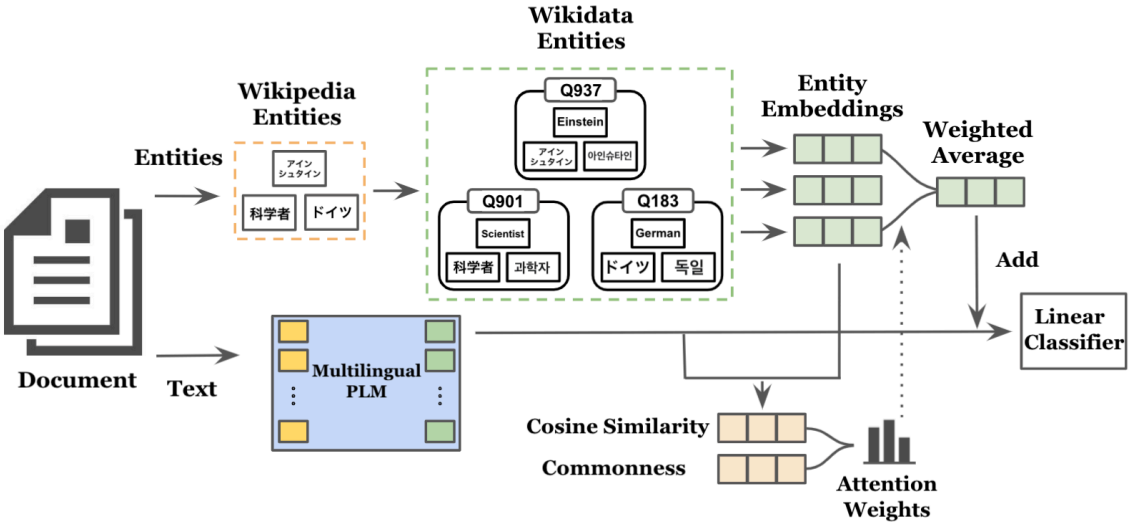


図 3.1: 多言語エンティティモデルの概要図。多言語エンティティモデルは文書から Wikipedia エンティティを抽出し、それらを対応する Wikidata エンティティに変換する。その後に注意機構に基づきエンティティベースの文書表現を計算する。この文書表現と、汎用多言語モデルから得られるテキストベースの文書表現を足し合わせた表現を言語間文書分類の入力特徴量として利用する。

3.3　提案手法

図 3.1 に提案手法である多言語エンティティモデルの概要図を表す。従来、汎用言語モデルではテキストを汎用言語モデルに入力し、得られるテキストベースの文書表現を特徴量としてテキスト分類タスクを解く。それに対し多言語エンティティモデルでは、文書が与えられた際、エンティティを抽出しそれらを全て Wikidata エンティティに変換する。それらの Wikidata エンティティからエンティティベースのテキスト表現を後述する注意機構構造に基づき計算し、テキストベースの文書表現に足し合わせる。得られた埋め込み表現を線形分類器に入力することでテキスト分類タスクを解く。

3.3.1　エンティティの検知

まず、多言語エンティティモデルでは入力文書からエンティティを検知するために、事前に知識ベースから以下の二つの辞書を構築する。

- エンティティ名-エンティティ辞書（エンティティ名と知識ベースエンティティを紐付ける辞書）
- 言語間リンク辞書（多言語のエンティティを対応する Wikidata エンティティに紐づける辞書）

ここでそれぞれの辞書の例を図 3.2 に示す。エンティティ名-エンティティ辞書は Wikipedia ページ内リンク情報 [51] を活用し、ハイパーリンクをエンティティ名、そのリンク先ページをエンティ

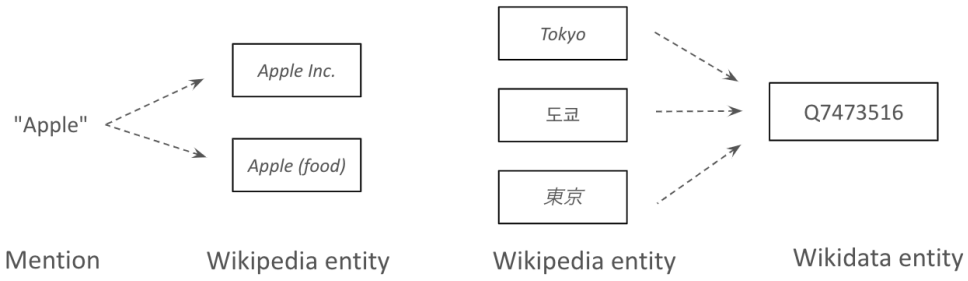


図 3.2: エンティティ名-エンティティ辞書（左）と言語間リンク辞書（右）の例

ティとして抽出することで構築する。 言語間リンク辞書は Wikidata における言語間リンク情報から構築する。

実際の学習・推論時には、まず文書中に登場する全ての単語・フレーズを抽出する。それらがエンティティ名-エンティティ辞書におけるエンティティ名のいずれかに一致する場合に、対応するエンティティを全て抽出する。ここで、エンティティ名-エンティティ辞書はエンティティ名に対応し得る全てのエンティティを抽出する（e.g., “東大” から “東京大学”、“東海大学”、“東大王”）ため、実際に対応していないエンティティを抽出してしまう場合があるが、これらの曖昧性解消はこの時点では行わず、後述する注意機構にて対処する。

次に、抽出されたエンティティ群を言語間リンク辞書を用いて Wikidata のエンティティに変換する。さらに、山田らの研究に倣いあるエンティティ名がハイパーリンクである確率（リンク確率）とあるエンティティ名が知識ベースに紐づく特定のエンティティを示す確率（コモンネス） [52] も事前に集計する。

3.3.2　モデル

次に、エンティティベースの文書表現を取得するため抽出した Wikidata エンティティそれぞれに対して事前学習エンティティ表現 $\boldsymbol{v}_{e_i} \in \mathbb{R}^d$ を割り当てる。ここで上述した通り本手法ではエンティティの曖昧性解消を行っておらず、実際には与えられた文書には関係していない Wikidata エンティティを抽出してしまっている場合がある。この問題を解決するために山田ら、Peters らの研究 [26, 34] に基づき文書に関連するエンティティを優先するためのエンティティ注意機構を導入する。実際には K 個の Wikidata エンティティが検出されたある文書 D に対して、以下の式に従いエンティティ表現 $\boldsymbol{v}_e$ の重みつき和としてエンティティベースの文書表現 $\boldsymbol{z} \in \mathbb{R}^d$ を計算する。

$$\boldsymbol{z} = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v}_{e_i}, \tag{3.1}$$

ここで $a_{e_i} \in \mathbb{R}$ はエンティティ e_i に対応する注意機構の重みを表し、以下の式によって学習される。

$$\boldsymbol{a} = \mathrm{softmax}(\boldsymbol{W}_a^\top \boldsymbol{\phi}), \tag{3.2}$$

$$\phi(e_i, d) = \begin{bmatrix} \mathrm{cosine}(\boldsymbol{h}, \boldsymbol{v}_{e_i}) \\ p_{e_i} \end{bmatrix} \tag{3.3}$$

ここで $\boldsymbol{a} = [a_{e_1}, a_{e_2}, \cdots, a_{e_K}]$ は注意機構の重みを表し、$\boldsymbol{W}_a \in \mathbb{R}^2$ は重みベクトルを表す。また、$\boldsymbol{\phi} = [\phi(e_1, d), \phi(e_2, d), \cdots, \phi(e_K, d)] \in \mathbb{R}^{2\times K}$ は文書 D から抽出された Wikidata エンティティ群それぞれと文書 D の関連度合いを測る素性であり、$\phi(e_i, d)$ はコモンネス p_{e_i} と汎用言語モデルから得られるテキストベースの文書表現 $\boldsymbol{h} \in \mathbb{R}^d$（e.g., mBERT の [CLS] トークンに対応する最終層の表現）とエンティティ表現 $\boldsymbol{v}_{e_i}$ のコサイン類似度を連結させたベクトルである。

そしてエンティティベースの文章表現 $\boldsymbol{z}$ とテキストベースの文書表現 $\boldsymbol{h}$ を足し合わせた表現をラベル c の確率を推論する分類器に入力する。

$$p(c \mid \boldsymbol{h}, \boldsymbol{z}) = Classifier(\boldsymbol{h} + \boldsymbol{z}). \tag{3.4}$$

なお、予備実験の際にエンティティベースの文章表現とテキストベースの文書表現を連結させた場合も検証しているが、足し合わせた方がゼロショット言語間文書分類タスクにおける提案モデルの性能が優れていたため、足し合わせる手法を採用した。

3.4 実験設定

本節では提案手法である多言語エンティティモデルの評価のために行った言語間文書分類タスクの実験設定について述べる。

3.4.1 データ

本実験では提案モデルを三つの言語間文書分類データセット（MLDoc [35]、TED-CLDC [36]、SHINRA2020-ML [37]）で評価した。本項ではそれぞれのデータセットについて述べる。

MLDoc

MLDoc [35] はニュース記事を CCAT (Corporate/Industrial)、ECAT (Economics)、GCAT (Government/Social)、MCAT (Markets) の 4 つのクラスに分類するシングルラベル文書分類タスクであり、8 言語での文書データが収録されている。実験では english.train.1000 を学習データとして用い、english.train.1000 を検証データとして用いた。また、MLDoc で多言語モデルの評価を行っている既存の研究 [35, 53] に倣い、正解率をこのタスクの評価指標として用いた。

Dataset	Language	Train	Dev.	Test
MLDoc	8	1,000	1,000	4,000
TED-CLDC	12	936	105	51–106
SHINRA	30	417,387	21,967	30k–920k

表 3.1: MLDoc, TED-CLDC, SHINRA2020-ML の 3 つのデータセットにおける統計量。

TED-CLDC

TED-CLDC [36] は TED [7]におけるトーク文書に対して 15 のトピックラベルを付与するマルチラベル文書分類タスクであり、12 言語での文書データが収録されている。このタスクは MLDoc と同様にトピック文書分類タスクであるが、より口語的な表現が多く、学習データが少ないため、より分類が困難なデータセットである。このデータセットにおける評価指標は既存の研究 [36] に倣い、F1 値のマイクロ平均を用いた。

SHINRA2020-ML

SHINRA2020-ML [37] は約 220 種類の拡張固有表現[8]（e.g., Person, Country, Government）を Wikipedia のページに付与するエンティティの型推定タスクであり、本実験ではマルチラベル分類タスクとして扱う。公開されている 30 言語の中で英語以外をテストデータとして利用した。このタスクでは SHINRA2020-ML での評価指標 [37] に倣い、F1 値のマイクロ平均を評価指標として用いた。

TED-CLDC と SHINRA2020-ML ではそれぞれ検証データとして学習データの 5% をランダムに抽出したものを利用した。全ての実験では英語を学習データを利用する原言語とし、その他の言語を目的言語とした。以上のデータセットの統計量を表 3.1 に示す。

3.4.2　エンティティの前処理

2019 年 1 月版の英語 Wikipedia dump[9]からエンティティ名-エンティティ辞書を構築した。また、2020 年 3 月版の Wikidata dump[10]から言語間リンク辞書を構築した。この Wikidata dump は 45,412,720 個の Wikidata エンティティを含む。この際、リンク確率が 0.01 以上かつコモンネスが 0.05 以上のエンティティを抽出した.

3.4.3　ベースラインモデル

本手法は任意の汎用多言語モデルに適用可能であるが、本実験における多言語エンティティモデルのベースモデルとしては mBERT [8] と XLM-R$_{base}$ [20] を用いた。huggingface らの既存実装[11]

[7] https://www.ted.com/talks?language=en
[8] https://ene-project.info
[9] https://dumps.wikimedia.org/enwiki/
[10] https://dumps.wikimedia.org/wikidatawiki/entities/
[11] https://github.com/huggingface/transformers

Model	MLDoc	TED-CLDC	SHINRA2020-ML
mBERT	32 / 2e-05	16 / 2e-05	128 / 5e-05
XLM-R	32 / 2e-05	16 / 5e-05	64 / 2e-05
M-BoE (mBERT)	32 / 2e-05	16 / 2e-05	128 / 5e-05
M-BoE (XLM-R)	32 / 2e-05	16 / 5e-05	64 / 2e-05

表 3.2: 実験で利用したハイパーパラメータ。それぞれの枠において左がバッチサイズ、右が学習率を表す。

を拡張することで多言語エンティティモデルを実装した。また、テキストベースの文書表現として [CLS] トークンに対応する最終層の隠れ表現 $\boldsymbol{h}$ を用いた。

シングルラベル分類タスクである MLDoc では汎用言語モデルのベースモデルに対して全結合層と Softmax 層を接続し Cross-entropy 損失を最適化することで学習した。また、マルチラベル分類タスクである TED-CLDC と SHINRA-2020ML では汎用言語モデルのベースモデルに対して全結合層と Sigmoid 層を接続し Binary-cross-entropy 損失を最適化することで学習した。マルチラベル分類タスクの推論時は Sigmoid 層の出力が 0.5 以上になるラベルを予測ラベルに追加した。

また、MLDoc におけるトピック分類タスクでは既存手法の言語間文書分類の state of the art モデルである、93 言語の対訳コーパスを用いて BiLSTM モデルにより構築された多言語文表現 LASER [54] と畳み込みニューラルネットワーク文書分類モデルに基づく多言語埋め込み表現を活用した MultiCCA [35] と比較した。ここで平等な資源での比較を行うため、英語の教師データセット以外に目的言語ごとに追加の学習データを利用する手法や対訳コーパスを利用する手法とは比較していない。

さらに SHINRA-2020ML を用いたエンティティの型推定タスクに関しては、3.3.1 節で述べたエンティティの検出法の代わりに、Wikipedia に存在するハイパーリンクから直接エンティティを抽出する手法の評価も行った。このモデルに対しても上述の注意機構や事前学習済みエンティティ表現を導入している。

3.4.4　詳細設定

本実験で用いたハイパーパラメータを表 3.2 に示した。これらのハイパーパラメータは英語の検証データで調整した。パラメータ更新の最適化アルゴリズムは AdamW [55] を用い、gradient clipping は 1.0 に設定した。

全ての実験では学習モデルが英語の検証データにおける評価指標の値が収束するまで学習を行なった。学習は異なる乱数シードで 10 回行い、その結果の平均と 95%信頼区間を表に示す。

3.5　実験結果

表 3.3, 3.4, 3.5 にトピック分類タスクとエンティティの型推定タスクの実験結果を示す。3 つのタスクにおける全ての目的言語で多言語エンティティモデル（M-BoE）はそれぞれのベースラインモ

Model	en	fr	de	ja	zh	it	ru	es	target avg.
MultiCCA [35]	92.2	72.4	81.2	67.6	74.7	69.4	60.8	72.5	71.2
LASER [54]	89.9	78.0	84.8	60.3	71.9	69.4	67.8	77.3	72.8
mBERT	94.0	79.4	75.1	69.3	68.0	67.1	65.3	75.2	71.4 ± 1.4
XLM-R	94.4	84.9	86.7	78.5	85.2	73.4	71.3	81.5	80.2 ± 0.5
M-BoE (mBERT)	94.1	84.0	76.9	71.1	72.2	70.0	68.9	75.5	74.1 ± 0.7
M-BoE (XLM-R)	**94.6**	**86.4**	**88.9**	**80.0**	**87.4**	**75.6**	**73.7**	**83.2**	**82.2 ± 0.6**

表 3.3: MLDoc データセットにおけるトピック分類タスクの正解率。“target avg.” は目的言語における結果の平均を表す。

Model	en	fr	de	it	ru	es	ar	tr	nl	pt	pl	ro	target avg.
mBERT	51.6	47.7	43.9	50.6	47.9	53.1	41.3	44.2	49.4	46.2	45.1	45.4	47.1 ± 1.4
XLM-R	51.5	49.5	49.7	48.7	48.3	51.2	45.6	51.3	48.8	46.3	48.3	48.4	49.1 ± 1.8
M-BoE (mBERT)	**52.9**	49.5	46.2	**53.3**	49.2	**54.7**	44.7	49.1	51.0	47.6	47.7	48.2	49.6 ± 1.1
M-BoE (XLM-R)	51.7	**50.0**	**53.8**	51.3	**52.3**	52.9	**50.5**	**53.1**	**52.0**	**49.3**	**50.5**	**49.6**	**51.8 ± 0.9**

表 3.4: TED-CLDC データセットにおけるトピック分類タスクの F 値。

デル（mBERT、XLM-R）を上回る性能を示した。また、対応のあるサンプルの t 検定で提案モデルの目的言語の平均スコアがベースラインモデルと比較し有意水準 0.05 で統計的有意性があることを確認した。特に mBERT を拡張した多言語エンティティモデルは顕著な改善が見られ、MLDoc での正解率は 2.7%、TED-CLDC での F1 値は 2.5%、SHINRA2020-ML での F1 値は 2.1%の性能向上が確認された。

さらに、興味深いことに提案手法における単純な辞書に基づきエンティティを抽出するモデルが Wikipedia ページに登場するエンティティ群を直接抽出するモデル（表 3.5: M-BoE（Oracle））とエンティティの型推定タスクで同程度の性能を発揮しており、これは提案手法の注意機構に基づいたエンティティ検知が有効であることを示唆している。

3.6 分析

多言語エンティティモデルについて理解を深めるため、MLDoc データセットを用いていくつかの検証を行った（表 3.6）。最初に注意機構や事前学習済みエンティティ表現、エンティティ検出法など多言語エンティティモデルの各コンポーネントの性能に対する影響を分析した（3.6.1 項、3.6.2 項、3.6.3 項）。次に言語ごとに検知されるエンティティ数の違いが多言語エンティティモデルの性能にどう影響するかを調べた（3.6.4 項）。最後に多言語エンティティモデルの注意機構によって選別された重要なエンティティを可視化することで定性的な分析を行った（3.6.5 項）。

3.6.1 注意機構の影響

多言語エンティティモデルにおける注意機構とその学習のために利用しているコサイン類似度とコモンネスの二つの素性が有効に機能しているかを確認するため、注意機構を取り除いたモデル

	fr	de	ja	zh	it	ru	es	ar	tr	nl	pt	pl	ro	hi	no
mBERT	68.5	84.2	81.3	80.7	85.2	81.4	85.6	57.4	50.7	55.6	80.4	77.7	76.9	81.8	83.6
XLM-R	73.0	82.6	77.4	75.1	84.2	81.0	85.3	58.9	69.1	63.7	79.8	80.0	76.9	83.3	82.4
M-BoE (mBERT)	69.3	**85.1**	**82.5**	**82.2**	86.4	83.2	**86.6**	61.9	54.0	59.0	81.7	79.4	**80.5**	82.9	84.8
M-BoE (XLM-R)	**77.4**	84.5	79.0	77.0	85.6	83.2	85.8	**63.3**	**72.3**	65.5	80.7	**81.8**	77.8	84.8	84.0
Oracle M-BoE (mBERT)	75.4	85.2	81.9	81.8	86.5	83.0	86.5	61.9	53.7	61.7	**81.8**	79.7	79.9	83.0	**84.8**
Oracle M-BoE (XLM-R)	76.5	84.8	79.6	77.2	85.5	**83.4**	86.2	63.0	71.8	**67.6**	80.4	81.5	78.8	**84.8**	83.2
	th	ca	da	fa	id	sv	vi	bg	cs	fi	he	hu	ko	uk	target avg.
mBERT	84.0	81.5	80.1	80.2	72.4	79.4	79.3	74.0	74.6	75.7	74.0	77.1	81.3	78.0	76.6 ± 0.7
XLM-R	81.4	79.0	81.0	82.4	75.5	75.5	80.7	76.0	77.9	74.7	70.5	73.1	82.6	74.3	77.1 ± 1.2
M-BoE (mBERT)	85.1	83.2	81.4	82.1	75.4	**82.4**	81.2	76.1	76.8	**77.6**	**78.1**	**79.2**	82.9	**80.0**	78.7 ± 0.5
M-BoE (XLM-R)	82.1	80.9	**83.3**	**84.1**	78.2	78.7	81.9	79.1	79.6	76.9	71.9	75.5	**84.0**	77.0	**79.2 ± 0.9**
Oracle M-BoE (mBERT)	**85.3**	**83.2**	82.3	82.4	75.5	82.0	81.6	76.6	77.4	77.4	77.8	78.7	83.3	79.9	79.0 ± 0.5
Oracle M-BoE (XLM-R)	81.8	81.2	82.9	83.9	**78.3**	78.2	**82.5**	**79.1**	**79.9**	77.1	71.8	75.8	83.92	76.9	**79.2 ± 0.9**

表 3.5: SHINRA2020-ML データセットにおけるエンティティ型推定タスクの F 値。

Setting	M-BoE (mBERT) target avg.	M-BoE (XLM-R) target avg.
Full model	**74.1**	**82.2**
Attention mechanism:		
without attention	70.5	81.1
commonness only	72.4	81.8
cosine only	72.8	81.8
Entity embeddings:		
random vectors	73.0	80.9
KG embedding	73.2	81.4
Entity detection method:		
entity linking	71.7	80.5
entity linking + att	73.0	81.9

表 3.6: MLDoc における提案モデルの分析結果

(表 3.6: without attention)、コモンネスのみで注意機構を学習したモデル（表 3.6: commonness only)、コサイン類似度のみで注意機構を学習したモデル（表 3.6: cosine only）のそれぞれの性能と比較した。

実験結果から注意機構や各素性を取り除いた場合は言語間文書分類の性能が劣ることが確認され、特に注意機構を削除した場合は mBERT では 3.6%、XLM-R では 1.1% 性能が低下している。この結果は注意機構やその学習に用いた各素性が有効に機能していることを示唆している。

3.6.2　エンティティ表現の影響

多言語エンティティモデルにおける事前学習済みエンティティ表現の影響を調査するため、Wikipedia2Vec によるエンティティ表現それぞれを（1）ランダムなベクトル、（2）知識グラフ表現 (Table 3.6: **Entity embeddings**) に置き換えたモデルと性能を比較した。ここで知識グラフ表現とは知識ベー

Model	en (train)	fr	de	ja	zh	it	ru	es	avg.
External entity linking	20.0	19.2	14.6	8.15	5.2	11.7	12.7	13.8	13.2
Dictionary-based method (ours)	105.8	97.8	78.9	47.9	34.5	53.2	64.6	72.3	64.2

表 3.7: MLDoc データセットにおけるエンティティ検出数の比較

スにおけるエンティティ群とそれらの関係性からなる知識グラフが低次元ベクトルで表現されたものである。

実験では知識グラフ表現として state of the art モデルの一つである ComplEx [56] を利用した。ComplEx による知識グラフ表現は wikidata5m データセット [57] にて `kge tool`[12]を用いて学習した。ベクトルの次元は Wikipedia2Vec に合わせて 768 次元にし、その他のハイパーパラメータはツールの wikidata5m-complex configuration で利用されているデフォルトの値に設定した。

実験結果は Wikipedia2Vec でエンティティ表現を初期化する場合が最も良い性能を発揮することを示している。また、知識グラフ表現で初期化する場合はランダムベクトルよりは良い性能であることを示している。

3.6.3　エンティティ検出手法の影響

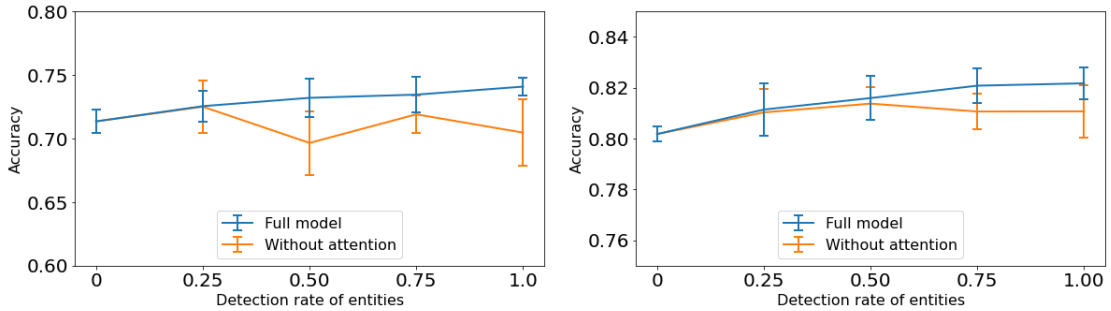
提案手法における辞書ベースのエンティティ検出手法の有効性を検証するため、多言語エンティティリンキングシステムによるエンティティ検出手法との比較を行った (表 3.6: **Entity detection method**)。このような多言語エンティティリンキングシステムは曖昧性のない知識ベースエンティティを検出する点で提案手法と異なる。多言語エンティティリンキングシステムとしては Google Cloud Natural Language API[13]を用い、検出した全てのエンティティを項 3.3.1 で述べた方法と同様に Wikidata エンティティに変換した。

実験結果から多言語エンティティリンキングシステムを用いた手法より提案手法による辞書ベースのエンティティ検出手法の方が性能が優れていることが確認された。この結果の原因として提案手法のエンティティの検出数の多さが考えられる。表 3.7 に示すように多言語エンティティリンキングシステムによるエンティティ検知数は提案手法と比較し、著しく少ない。これは多言語エンティティリンキングシステムは提案手法とは異なり曖昧性のないエンティティのみを検出し、文章中のエンティティ名ではないがエンティティとなり得る語（non-named entity）を検出しないためである。

従って提案手法が多言語エンティティリンキングシステムより優れている理由は（1）non-named entity も重要な文書中の特徴であること（2）エンティティリンキングが曖昧性解消に成功しなかった場合、正しいエンティティが考慮されなくなる点にあると考えられる。

さらに実験結果の節 3.5 でも述べたように、提案手法によるエンティティ検出手法はエンティティの型推定タスクで Wikipedia のページ情報から直接エンティティを抽出する手法と比較しても同程度の性能を発揮している。

[12] (`https://github.com/uma-pi1/kge`)
[13] `https://cloud.google.com/natural-language`



(a) M-BoE (mBERT)　　(b) M-BoE (XLM-R)

図 3.3: MLDoc データセットにおけるエンティティ検出率ごとの文書分類の正解率

Dataset		Model	fr	de	it	ru	es	ja	zh	ar	tr	nl	pt	pl	ro	Pearson
MLDoc	#Ent		97.8	78.9	53.2	64.6	72.3	47.9	34.5	-	-	-	-	-	-	
	Rate	mBERT	5.8	2.4	4.3	5.5	0.4	2.6	6.2	-	-	-	-	-	-	-0.13
		XLM-R	1.8	2.5	3.0	3.4	2.1	1.9	2.6	-	-	-	-	-	-	-0.34
TED-CLDC	#Ent		218.9	223.5	217.8	227.2	227.9	-	-	227.3	185.0	190.7	166.4	134.5	211.2	
	Rate	mBERT	3.8	5.2	5.7	2.7	3.0	-	-	8.2	11.1	3.2	3.0	5.8	6.2	-0.11
		XLM-R	1.0	8.2	5.3	8.3	3.3	-	-	10.7	3.5	6.6	6.5	4.6	2.5	0.17

表 3.8: 各目的言語におけるエンティティの検知数 (#Ent) と評価値の向上率 (Rate) のピアソンの相関係数

次に、Wikidata エンティティの検出数の性能への影響を調べた。具体的には多言語エンティティモデル（Full model）と注意機構を取り除いたモデル（Without attention）に対して学習時と推論時に検出されるエンティティをランダムに取り除き性能の推移を観測した。

図 3.3 が示す実験結果からエンティティ検出率が高ければ高いほど、提案モデルの性能が向上することが確認された。しかし注意機構を取り除いたモデルではエンティティ検出率が高くなってもモデルの性能は向上せず一貫した結果が得られなかった。これらの結果はエンティティの検出数が多ければ多いほど提案モデルの性能が向上し、その安定した性能向上には注意機構による重要なエンティティを選別する効果が寄与していることを示唆している。

3.6.4　言語の違いの影響

提案手法では推論時の Wikidata エンティティの検出数が目的言語によって異なるため、この検出数の違いが提案モデルの言語間文書分類の性能にどのように影響するかを調べた。具体的には MLDoc と TED-CLDC データセットにおけるそれぞれの目的言語にて（1）ベースラインモデルに対する性能の向上率、（2）1 文書あたりのエンティティ検知数の平均をそれぞれ計算し、それらのピアソンの相関係数を計算した。表 3.8 はその結果を表す。

実験結果では明らかな相関関係は見られなかった。これはすなわち推論時における言語ごとのエンティティ検出数の違いは多言語エンティティモデルの性能へあまり影響しないことを示唆する。

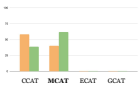
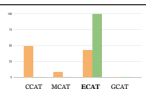
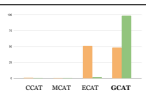
Language	Document	Label	Probability distribution M-BERT M-BoE	Top three entities
Ja	［台北 2日 ロイター］引け前の台湾株式市場で、加権指数が3.28％急落した。フローカーらによると、工業株に売りが集中したため、という。大引け前10分(0350gmt)現在、加権指数は278. 07ポイント(3. 28％)急落し、8207. 59。売買代金は、1090億台湾ドル。	**MCAT** (Markets)		**"Stock certificate"** **"Share price"** **"Taiwan Capitalization Weighted Stock Index"**
Zh	〔路透社東京19日電〕日本大藏省一顧問小組週四促請大藏省取消目前只允許被授權外匯銀行進行外匯交易的管制,完全開放外匯市場交易資格的限制. 這項限制的取消將使投資人進出外匯市場更為容易;此外,銀行業也可藉此增進競爭力,並促進市場的流動性及活絡匯市的交易.(完)	**ECAT** (Economics)		**"Ministry of the Treasury"** **"Financial transaction"** **"Competition (economics)"**
Ru	москва, 17 мар (рейтер) - президент рф борис ельцин подписал федеральные законы о внесении изменении и дополнении в статьи 100 и 110 закона рф "о государственных пенсиях в рф", сообщила пресс-служба президента рф. статья 100 закона излагается в следующеи редакции: "в заработок для исчисления пенсии включаются все виды выплат (дохода), полученных в связи с выполнением работы, предусмотреннои статьеи 89 закона, на которые начисляются страховые взносы в пенсионный фонд рф". пресс-служба президента рф сообщила, что виды выплат, на которые не начисляются страховые взносы в пенсионный фонд рф, определяются правительством рф.	**GCAT** (Government Social)		**"Federal law"** **"Pension Fund of the Russian Federation"** **"Kremlin Press Secretary"**

図 3.4: MLDoc データセットにおける実験結果の例。“Top three entities” は注意機構によって選別された最も影響のある（最も注意機構の重みの大きい）3 つのエンティティを表す。

これは提案手法における辞書ベースのエンティティ検出手法が比較的検知数の少ない言語であってもエンティティの特徴をとらえるのに十分な量のエンティティを検出しているためだと考えられる。例えば MLDoc データセットで最も検知数の少ない中国語であっても平均 34.5 個のエンティティが検出されており、性能向上率が他の言語と同等以上であることが実験結果からわかる。

3.6.5　定性評価

さらに多言語エンティティモデルがどのようにして性能改善を達成したかについて分析するため、定性的な分析を行った。具体的には MLDoc データセットにおいて mBERT は分類に失敗したが、多言語エンティティモデルは正しい分類を行った文書のうち、エンティティが顕著に影響していた（注意機構の重みが大きかった）例を 3 つ表 3.4 に示した。3 つの例における多言語エンティティモデルは文書に関連するエンティティを重要視している。

例えば台湾株式市場における日本語の文書では、多言語エンティティモデルの注意機構は ***Stock certificate***, ***Share price***, and ***Taiwan Capitalization Weighted Stock Index*** を重要視しており、これらは文書のトピックを捉えたエンティティであることがわかる。

3.7　結論

3.7.1　まとめ

本研究では、同じ概念を表すエンティティが言語に依存せず定義されている Wikidata 知識ベースを用いて多言語汎用言語モデルを拡張することで、ゼロショット言語間文書分類の性能を向上させる多言語エンティティモデルを提案した。多言語エンティティモデルは文書から Wikidata エンティティを抽出し、それらのエンティティ表現から文書への関連度を考慮した重みつき和を計算することでエンティティベースの文書表現を計算する。この表現をテキストベースの文書表現に足し合わせた特徴量を用いてゼロショット言語間文書分類タスクを解く。文書のトピックを効率良く捉えるエンティティ表現が多言語で共有されるため、原言語のエンティティで学習した多言語モデル

は複数の目的言語における推論時においても直接エンティティの特徴量を考慮することが可能となる。

実験では 3 つのゼロショット言語間文書分類タスクでベースラインとなる汎用言語モデルや既存の state of the art モデルに勝る性能を発揮した。さらに多言語エンティティモデルの各コンポーネントの性能への影響や、検出されるエンティティ数の影響、注意機構により重要視されているエンティティの可視化を行うことで多言語エンティティモデルに対する知見を深めた。

3.7.2 今後の展望

本項では多言語エンティティモデルの今後の展望について述べる。

一点目は本手法を他の種類の自然言語処理タスクで評価することである。本実験ではゼロショット言語間転移タスクとして文書分類タスクを用いているが、ゼロショット言語間転移タスクには XNLI や Cross-lingual Question Answering など様々な自然言語処理タスクが存在する。よって本手法のような（1）文書のトピックを捉え、（2）異なる言語の橋渡しを行う効果を持つ Wikidata エンティティが他のどのようなタスクに応用可能であるかを検証することは本手法の汎用性の高さを確かめる上で重要な知見となると考えられる。

二点目は他の多言語汎用言語モデルと組み合わせることである。本実験では多言語エンティティモデルを M-BERT と XLM-R の 2 つのみをベースラインモデルとして用いているが、これらの手法はどちらも BERT に基づいたモデルであり、似たような振る舞いをすることが期待される。しかし多言語モデルには LSTM に基づくモデル [54] や多言語埋め込み表現に基づくモデル [18] など様々なモデルが存在するため、これらのモデルに対しても本手法のようなエンティティによるモデルの拡張の有効性を確かめることが考えられる。

三点目は目的言語の検証データを用いてモデルのエポック数を調整した場合の性能の検証である。本実験では多言語モデルの学習の収束条件として英語の検証データにおける性能が最も高いエポック時のモデルとしている。しかし Keung ら [53] は原言語の検証データにおける性能と目的言語における推論時の性能は必ずしも明確な相関があるわけではないことを実験で示しており、目的言語の検証データで調整した結果を多言語モデルのゼロショット言語間転移の性能限界の目安として掲載することを推薦している。この研究に基づき、目的言語ごとの検証データでエポック数を調整したモデルの性能を検証することが考えられる。

第4章　エンティティを利用したContrastive learningによる多言語文表現の学習

4.1　序論

文表現とは、文をその文の持つ意味を表す数値ベクトル表現で表したものである。文の数値ベクトル化により言語をコンピュータが扱うことが可能となり、 類似文検索や文書クラスタリング、パラフレーズ検知など、文を入力とする様々な自然言語処理タスクの処理が容易になる。従来、この文表現を生成する手法として bag-of-words に基づくカウントベースの手法が提案されてきたが [58]、これらの手法は周りの文の文脈や文中の単語同士の関係性が捉えられず、文の意味を十分に埋め込めているとは言い難い。これらの手法に対し近年では、Skip-Thought と呼ばれる Skip-gram [14] における分布仮説を文レベルに拡張した手法を皮切りに、文をベクトル表現として直接エンコードする汎用文表現モデルの研究が盛んに行われており、Semantic Textual Similarity (STS) [59] などのタスクで目覚ましい成果が報告されている。

文表現モデルの中でも、入力文の言語に依存せず文をベクトルへエンコードできるモデルは多言語文表現モデルと呼ばれる。これらのモデルの多くは、対訳コーパスや XNLI など言語間で意味的に関連する文が対応付けされたデータ （言語間資源）を用いて原言語と目的言語の文表現を近づける教師あり学習を行う [60, 61]。しかしこれらの言語間資源は多言語話者によるアノテーション作業によって構築されるものであり、言語によっては容易に手に入らない場合がある。従って、既存の教師あり多言語文表現モデルは言語間資源が豊富に存在する言語群における学習しか行えないという制限がある。

このような問題の緩和のため、本研究では、知識ベースの言語資源を活用した文表現学習手法について探求する。知識ベースは幅広い言語で容易に手に入り、言語に非依存であるためいくつかの研究で多言語モデルの改善に活用されている [62, 63]。知識ベースを多言語文表現学習に活用することで、既存の手法で必要とされていた対訳コーパスに依存しない幅広い言語で機能する多言語モデルが構築されることが期待される。

本研究で提案する文表現学習手法 Entity-Aware Contrastive Learning of Sentence Embedding (EASE) では、Wikipedia におけるハイパーリンク情報から文とエンティティのペアデータを構築し、このデータを用いて文と関連するエンティティを近づけ、関連しないエンティティを遠ざける Contrastive learning を行うことで文表現を学習する (図 4.1)。この学習により、言語に非依存な知識ベースエンティティを軸に様々な言語の文が共有ベクトル空間に埋め込まれるため、言語に非依存な文表現が学習されることが期待される。また、EASE では似た意味を持つ複数の文が共通のエンティティを軸に類似度の高いベクトルとして埋め込まれる効果も期待される。

実験では、まず EASE を単言語のデータセットを用いて英語汎用言語モデルに適用し、STS と

Label	Sentence1	Sentence2
contradiction	A man inspects the uniform of a figure.	The man is sleeping.
	is sleeping.	
neutral	A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.
entailment	A soccer game with multiple males playing.	Some men are playing a sport.

表 4.1: NLI データセットの例

Short Text Clustering (STC) タスクで既存の state of the art の教師なし文表現モデルと比較して勝るもしくは同等の性能を発揮することを確認した。次に、EASE を複数言語のデータセットを用いて汎用多言語モデルに適用し、多言語 STS、多言語 STC、対訳コーパスマッチング、言語間文書分類タスクなどの幅広いタスクで既存の多言語文表現モデルより優れた性能を発揮した。さらに、大規模な対訳コーパスで学習済みの既存の多言語文表現モデルに対して、EASE でファインチューニングを行うことで低資源言語における性能を補完できることを示した。

4.2 関連研究

4.2.1 文表現

文表現を生成する古くからの手法として bag-of-words に基づくカウントベースの手法がある [58]。しかしこれらの手法は文の周りの文脈や文中の単語間の関連性を捉えることができておらず、文の意味を十分に埋め込めているとは言えない。これらの手法に対し、 Kiros ら [64] は直接文をベクトル表現にエンコードする手法として Skip-gram を文レベルに拡張した Skip-Thought を提案した。Skip-Thought ではエンコーダ、デコーダに gated recurrent unit (GRU) [65] を用い、文書中の i 番目の文 S_i からその前後の文 S_{i-1}, S_{i+1} を出力させるような系列変換タスクを学習させる。その際にエンコーダ側から得られる隠れ層の表現を Skip-Thought 文表現として用いる。 Skip-Thought は教師なし学習であり、アノテーションされたデータを一切必要とせず学習可能な手法である。また、InferSent [66] では Natural Language Inference (NLI) データセットを活用した、教師ありの設定で文表現を学習している。NLI データセットとは自然言語推論を学習するためのアノテーションがなされたコーパスであり、表 4.1 のように 2 つの文とその文の関連性を表す entailment、contradiction、neutral のいずれかのラベルが付与されている。InferSent では BiLSTM エンコーダから max pooling を行うことで得られる表現を文表現とし、NLI データセットの文の関係ラベルを活用した距離学習を行う手法を提案している。

これらの研究に対し、近年では事前学習済みの汎用言語モデルを活用することで文表現を学習する手法が提案されている。単純な手法としては事前学習済み汎用言語モデル BERT から得られる [CLS] トークンに対応する表現や最終層の各トークンに対応する表現の平均を文表現として用いることが考えられるが、これらは表現同士の意味的類似度を計算する場合には適しておらず、単語分散表現の平均などよりも後述する Semantic Textual Similarity (STS) などの文表現評価タスクにおける性能が悪い場合が多い [67]。以上の背景から多くの研究では汎用言語モデルに対して何らかの形で fine-tuning を行うことで文表現を学習している。

代表的な手法である Sentence-BERT [67] では、NLI データセットや STS データセットを用い BERT や RoBERTa を文表現のエンコーダとして距離学習を行う。Sentence-BERT では学習データの種類（文同士の関係ラベルが付与されているか、文同士の類似度スコアが付与されているか）に合わせて 3 つの学習手法を提案している。その 1 つの Classification Objective では以下の式で表されるような 2 文の文表現をそれぞれ u, v とするとそれぞれ u と v, $|u - v|$ を連結したベクトルを Softmax 分類器に入力し、文同士の関係ラベルを分類するタスクを解く。

$$L = \mathrm{softmax}(W_t(u, v, |u - v|))$$

ここで $W_t \in \mathbb{R}^{3n \times k}$ は全結合層である。

Regression Objective では u と v のコサイン類似度をとり文の類似度スコアとの平均二乗誤差を最適化する。Triplet Objective はあるアンカー文の文表現 s_a に対して正例 s_p には近づけ、負例 s_n には遠ざける以下のような目的関数を最適化する。

$$L = max(||s_a - s_p|| - ||s_a - s_n|| + 1, 0)$$

Sentence-BERT は後述する STS タスクや、文表現を別の分類器の特徴量として用い、文書分類などの下流タスクを解く SentEval [68] において発表当時の state of the art を達成しており、近年の BERT に基づく多くの文表現学習手法のベースラインモデルとして採用されている。

4.2.2　多言語文表現

また入力文の言語に依存せず、文表現を獲得できる多言語文表現モデルについても研究が進められている。この多言語文表現はゼロショット言語間転移学習や言語間の類似文検索、コンパラブルコーパスから対訳文を同定する bitext mining など様々な応用先がある。

多言語文表現を学習する多くの手法では大規模な対訳コーパスを用いた教師あり学習を行う。代表的な手法である LASER [69] は英語・フランス語を目的言語とする約 220 万ペアの対訳コーパスを学習データとして用い、エンコーダ・デコーダモデルにおける機械翻訳の学習と同様にして stacked BiLSTM を学習する。その際にエンコーダ側の出力を max-pooling したベクトルを多言語文表現として提案している。Reimers ら [70] は英語 Sentence-BERT を教師モデル、汎用多言語モデルを生徒モデルとして対訳コーパスを活用し知識蒸留を行うことで効率的に多言語文表現を学習する手法を提案している。LaBSE [60] では約 60 億ペアの対訳コーパスを活用し、以下の式で表される additive margin softmax loss with in-batch negative sampling [71] を活用した距離学習により BERT に基づくモデルを学習している。

$$L = -\frac{1}{N}\sum_{i=1}^{N} \frac{e^{\mathrm{sim(x_i, y_i) - m}}}{e^{\mathrm{sim(x_i, y_i) - m}} + \sum_{n=1, n\neq i}^{N} e^{\mathrm{sim(x_i, y_n)}}} \tag{4.1}$$

ここで x_i, y_i はそれぞれ原言語、目的言語の文表現を表し、N はバッチサイズ、$sim(\mathbf{h_1}, \mathbf{h_2})$ はコサイン類似度 $\frac{\mathbf{h_1}^\top \mathbf{h_2}}{\|\mathbf{h_1}\| \cdot \|\mathbf{h_2}\|}$ を表す。

DuEAM [61] では XNLI データセットの言語間ペアデータを活用することで原言語と目的言語における文表現の距離学習を行い、対訳コーパスを利用したモデルに迫る性能を達成している。

これらの手法は後述する多言語 STS や bitext mining などのタスクで目覚ましい性能を達成しているが、対訳コーパスや XNLI など限られた言語にしか存在しない言語資源に依存しており、低資源言語へは適用できないという問題点は残る。本研究では 300 言語以上をカバーする Wikipedia コーパスを活用し、多言語文表現を学習する手法を提案する。

4.2.3　エンティティを用いた文表現の学習

知識ベースに紐づくエンティティを用いた文表現の学習手法として、山田らの Neural Text-Entity Encoder (NTEE) [72] がある。NTEE では文書中に登場するエンティティがアノテーションされている DBpedia abstract corpus [73] を用い、文書からエンティティを推測するようなタスクを解くことで文書表現を学習する。NTEE は後述する STS や Entity linking などのエンティティ関連の自然言語処理タスクで高い性能が報告されている。

NTEE では単語分散表現やエンティティ表現を Skip-gram で学習した表現で初期化し、単語分散表現の和を文書表現としている。これに対し、提案手法では Transformer に基づく汎用言語モデルの出力表現を、事前に Wikipedia を用いて学習したエンティティ表現に近づけるような学習を Contrastive learning のフレームワークに従って行うことで文表現を学習する。

また、DOCENT [74] では提案手法と同様にエンティティとそれに関連する文を近づける学習を行う。具体的には 3 種類のエンティティを用いた事前学習手法（DUAL、FULL、HYBRID）を提案している。DUAL では文からエンティティを推測する。また、FULL ではエンティティトークンと文を連結させた文を用いて Masked LM タスクを解く学習を行い、HYBRID では FULL のように MLM タスクを解く際、BERT からの文埋め込み表現にエンティティ表現を連結させ、Masked LM タスクを解く。この手法は提案手法と類似した学習を行っているが、学習の目的がエンティティ表現の学習にある。提案手法はエンティティの言語非依存性を活用し、様々な自然言語処理タスクへの利用を想定した多言語文表現を学習することを目的としている。

4.2.4　Contrastive Learning による文表現の学習

Contrastive Learning とはあるデータ（アンカーデータ）に対して意味が近いデータ（正例）を近づけ、意味が異なるデータ（負例）を遠ざけることでベクトルを学習する表現学習の手法である [75]。画像処理分野においては SimCLR [76] と呼ばれる手法で、ある画像データからクロップや回転などのデータ拡張により生成したデータを正例として利用し、ランダムな画像データを負例として利用するような Contrastive learning を導入することで、自己教師あり学習や半教師あり学習における画像分類タスクで既存手法の性能を大幅に上回る state of the art の性能を達成し、注目を浴びた。

SimCLR では実際には i 番目のデータ x_i に対して以下のような目的関数に従い、Contrastive learning を行う。

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{h_i^+})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{h_j^+})/\tau}}, \tag{4.2}$$

STS score	Sentence1	Sentence2
2.6	Three men are playing chess.	Two men are playing chess.
5.0	A plane is taking off.	An air plane is taking off.

表 4.2: STS-B データセットにおける STS スコアの例。

ここで x_i^+ は x_i の正例を表し、$\mathbf{h_i}, \mathbf{h_i^+}$ はデータ x_i, x_i^+ から得られるベクトルを表す。また N はバッチサイズを表し、$sim(\mathbf{h_1}, \mathbf{h_2})$ はコサイン類似度 $\frac{\mathbf{h_1}^\top \mathbf{h_2}}{\|\mathbf{h_1}\| \cdot \|\mathbf{h_2}\|}$、$\tau$ は温度を表すハイパーパラメータである。

この目的関数は cross-entropy objective with in-batch negatives と呼ばれ [77]、ミニバッチデータの中で正例 x_i^+ 以外のデータ $(x_1^+, x_2^+, \cdots, x_{i-1}^+, x_{i+1}^+, \cdots, x_N^+)$ を負例とする目的関数である。近年の Contrastive Learning を用いた多くの研究ではこの目的関数に基づいた学習が盛んに研究されている。

自然言語処理の分野においても Contrastive Learning に基づいた文表現学習手法がいくつか提案されており、特にどのような正例ペアを用いるかが議論の中心になっている。ConSERT [78] では敵対的摂動やトークンのシャッフル、トークンやトークベクトル要素をランダムに取り除くカットオフなど様々なデータ拡張の手法により正例データを生成している。また、NLI データセットにおける entailment ラベルや contradiction ラベルの文ペアをそれぞれ正例ペア、負例ペアとして利用する手法も提案されている [79, 80]。

SimCSE、Mirror-BERT [80, 81] では、BERT に基づく汎用言語モデルの Transformer 層の内部にある dropout [82] のランダム性のみ異なるエンコーダから得られる二文の文表現を正例としている。特に SimCSE では STS や NLI データセットを一切利用しない教師なしの設定で発表当時の state of the art の性能を大幅に更新している。

4.2.5　文表現の評価

本節では文表現を評価するタスクについて解説する。

Semantic Textual Similarity

Semantic Textual Similarity (STS) タスク [59] とは二文が与えられた際にその文の意味的な等価性を評価するタスクである。このタスクでは二文の意味的な類似度として STS 値と呼ばれる 0（意味が全く異なる）から 5（意味が等価である）までの離散値が与えられる。このタスクを用いてモデルを評価する際は、モデルが出力した二文の文表現におけるコサイン類似度などの値と人間がアノテーションした STS 値のピアソンやスピアマンの相関係数を計測する。この相関係数が高いほど、より人間の直感に近い文の意味が表現に埋め込まれていると言える。

STS-Benchmark (STS-B) [59] の実際の例を表 4.2 に示す。一つ目の例における “Three men are playing chess.” と “Two men are playing chess.” とは “Three” と “Two” とが異なるのみであるが、STS 値は低く 2.6 となっている。それに対し二つ目の例では “A plane is taking off.” と “An air plane is taking off.” は “A plane” と “An air plane” が異なるが、この二つの単語の意味は全

く同じ意味であるため、STS 値は 5.0 となっている。このように STS での意味的類似度評価は人間が判断した値になっており、n-gram に基づいて計算されるような類似度計算手法とは性質が異なる。

また多言語 STS と呼ばれる、様々な言語の文ペアにおいて文の意味的な等価性を評価するタスクも提案されており、多言語文表現の評価指標の 1 つとして利用されている。

Short Text Clustering

多くの既存研究では STS に関連するタスクを解くことで文表現の有効性を示している。しかし STS タスクでは文表現が細かい意味的含意や矛盾を推論できるかを確認できるが、より大まかなカテゴリ構造情報を捉えているかは確認できないと指摘されている [79]。大まかなカテゴリ構造情報とはすなわち、文表現空間上で似たカテゴリに所属している文は近傍に埋め込まれ、逆に異なるカテゴリに所属している文は離れた空間に埋め込まれているような状態を表す。

例えば、近年の文表現学習手法で主に利用される NLI データセットの矛盾ペア（負例）の例として “A dog catching a Frisbee” と “A dog eating food” があるが、この 2 つの文は意味的に矛盾していると捉えられるが、より大まかな「犬が何らかの行動をしている」という観点では同じカテゴリに所属する文とも考えらえる。この観点は例えば文表現を文検索などの下流タスクに応用する上で重要であると考えられる。

大まかなカテゴリ構造情報を評価するタスクとして文をクラスタリングする Short Text Clustering (STC) が提案されている。Zhang らはトピック別にラベルが付与された文データから得られる文表現群を K-Means 法 [83] によってクラスタリングを行い、Hungarian アルゴリズム [84] によってクラスタリングの正解率を計算している。

また、この STC の性能（文表現が大まかなカテゴリ構造情報を捉えられるか）と上述の STS の性能（文表現が文の細かい意味的含意や矛盾を捉えられるか）はトレードオフの関係にあることが示されている [79]。よって両者のタスクで高性能な文表現を学習することは、汎用文表現を作成する上で重要な課題となっている。

対訳コーパスマッチング

多言語文表現を評価するタスクとして対訳コーパスマッチングがある。このタスクでは対訳コーパスのペアが複数与えられた際に、原言語の文に対応する目的言語の翻訳文の探索を行う。具体的には原言語のある文の文表現と目的言語の文それぞれに対してコサイン類似度を計測し、最も類似度が高い文が実際の翻訳文と一致しているかを計測する。このタスクのデータセットとしては Artetxe ら [69] の提案した Tatoeba があり、英語を原言語、112 の言語を目的言語としてそれぞれ 1000 文で構成されている。

言語間転移タスク

原言語の教師データで学習したモデルを他の目的言語に適用する言語間転移タスクは多言語文表現を評価する場合にも用いられる。代表的なデータセットとしては言語間文書分類を行う MLDoc

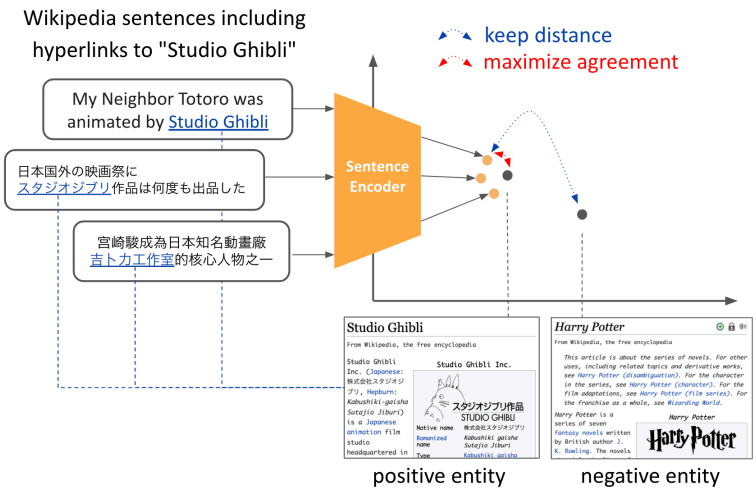


図 4.1: EASE におけるエンティティによる Contrastive learning のイメージ図。 EASE では Contrastive learning のフレームワークに従い、文表現がその文中に登場するハイパーリンクエンティティの表現に近づき、関連しないエンティティ表現とは遠ざかるようにモデルが学習される。ここでエンティティ表現は言語を跨ぎ共有されているため、文表現は言語に非依存となることが期待される。

(3.4.1 項を参照）や言語間で自然言語推論タスクを解く XNLI [22] がある。

Uniformity と Alignment

Contrastive learning による表現学習手法を評価する指標として alignment と uniformity が提案されている [85]。alignment は正例ペアのデータ群 p_{pos} が与えられた際に、以下の式により正例ペアの距離を計測する。

$$l_{\text{align}} = \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2. \tag{4.3}$$

また、uniformity は表現がどのくらい均等に分布しているかを計測する指標であり、以下の式によって表される。

$$l_{\text{uniform}} = \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}, \tag{4.4}$$

ここで p_{data} は与えられたデータの分布を表す。

4.3　提案手法

本節では、Wikipedia から得られる文書とそれらに紐づくハイパーリンクエンティティから文表現を学習する手法、EASE について述べる (図 4.1)。

Sentence	Entity
3 月 24 日 - 宮崎駿監督「千と千尋の神隠し」が 第 75 回アカデミー賞長編アニメ映画賞を受賞。	Q155653 (Spirited Away)
In 2003, after winning an Oscar for his film Spirited Away, Hayao Miyazaki received Le Guin's approval but was busy directing Howl's Moving Castle."	Q155653 (Spirited Away)

表 4.3: エンティティ-文データセットの例。

4.3.1　エンティティの処理

Wikipedia の各文においてハイパーリンクが存在する場合、その文とハイパーリンクに対応するエンティティをエンティティ-文ペアとして抽出する。それらの Wikipedia エンティティ（e.g. “東京大学”、 “University of Tokyo”）は西川ら [62] や李ら [63] の手法に倣い、言語間リンク辞書を用いて Wikidata のエンティティ（e.g. Q7842）に変換する。またこの際に後述するハードネガティブエンティティの構築のため、エンティティの型を Wikidata から取得する。実際に構築されたエンティティ-文データセットの例を表 4.3 に示す。

4.3.2　エンティティを用いた Contrastive learning

文とそれに関連するエンティティのペアデータ $\mathcal{D} = \{(s_i, e_i)\}_{i=1}^{m}$ が与えられた際、s_i に対応する文表現 $\mathbf{s_i} \in \mathbb{R}^{\mathrm{d_s}}$ からそのペアの e_i を表すエンティティ表現 $\mathbf{e_i} \in \mathbb{R}^{\mathrm{d_e}}$ を予測する学習を行う。実際には Chen ら [86] の Contrastive learning のフレームワークに従い以下のような目的関数を最適化する。

$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_j})/\tau}}, \tag{4.5}$$

ここで、$\mathbf{W} \in \mathbb{R}^{\mathrm{d_e} \times \mathrm{d_s}}$ は全結合層を表す。

この手法により意味的類似度の高い各言語の様々な文が、共通のエンティティベクトルの近傍に埋め込まれる。

ハードネガティブの導入

アンカーデータとは区別が難しいデータを負例として導入することで学習を困難にするハードネガティブの有効性がいくつかの研究で示されている [80, 87]。Wikipedia から得られるエンティティ情報をさらに活用するために以下の二つの条件を満たすエンティティをハードネガティブエンティティとして導入した。

- 正例エンティティと同じ型を持つ
- 正例エンティティと同じ Wikipedia のページ上に現れない

これらのハードネガティブエンティティ（e.g. “Ken Hirai”）は正例エンティティ（e.g. “Kyary Pamyu Pamyu”）と同じ型（e.g. “human”）であるため、エンティティ表現上では類似度の高いベクトルで表される。しかし同じ Wikipedia のページ上に現れないため実際には関連性が薄い場合があり、これらを正例エンティティと区別することでより細かい意味の違いを捉えた文表現が学習されることが期待される。ハードネガティブエンティティe^-を含むデータセット$\mathcal{D} = \{(s_i, e_i, e_i^-)\}_{i=1}^m$が与えられた際、ハードネガティブを導入した目的関数は以下のとおりである。

$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^N (e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_j})/\tau} + e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We^-_j})/\tau})}, \tag{4.6}$$

4.3.3 自己教師あり Contrastive learning

また、ある入力文の文表現から BERT に基づくモデルの Dropout ノイズのみが異なる文表現を予測するような自己教師あり Contrastive learning を行う学習の有効性が報告されており [81, 80]、提案手法ではこの学習を上記のエンティティを用いた Contrastive learning と組み合わせることで、より文の細かい意味を捉えた文表現の学習を目指す。目的関数は以下のようになる。

$$l_i^s = -\log \frac{e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{s_i^+})/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{s_j^+})/\tau}}, \tag{4.7}$$

以上のエンティティを用いた Contrastive learning と自己復元学習を組み合わせ、最終的には目的関数は以下のようになる。

$$l_i^{ease} = \lambda l_i^e + l_i^s \tag{4.8}$$

ここでl^e、l^sはそれぞれ（4.6）式、（4.7）式で定義した目的関数である。またλはエンティティを用いた Contrastive learning と自己教師あり Contrastive learning のバランスを定めるハイパーパラメータである。

4.4 実験

本節では提案手法である EASE の評価のために行った実験の設定について述べる。実験では複数言語の Wikipedia データセットを用いて学習する多言語設定と単言語の Wikipedia データセットを用いて学習する単言語設定の 2 つの設定で EASE の有効性を検証した。

4.4.1 ベースラインモデル

多言語の設定では汎用多言語モデルである M-BERT [8]、XLM-R [20] をベースモデルとし、これらの事前学習がなされたチェックポイントから提案手法による学習を行った。また Gao らの SimCSE モデルにおいて同じ複数の言語データセットで学習したモデルとも比較した。単言語の設定では汎用言語モデル BERT [8]、RoBERTa [88] をベースモデルとし、これに対して提案手法を適用さ

せた。さらに教師なしで文表現を学習する既存手法として単語分散表現の代表的なモデルである Glove embeddinng [89] や、近年の Contrastive learning に基づく state of the art の教師なし文表現モデルである SimCSE [80]、CT [90]、DeCLUTR [91] と比較した。

4.4.2 学習データ

実験では 2019 年 1 月版の Wikipedia を用い、`polyglot`[14] ライブラリにおける文トークナイザにより文章を文に分割した。分割した各文におけるハイパーリンクからエンティティを抽出する際はハイパーリンクとして 10 回以上 Wikipedia データに登場するエンティティに絞り、さらにそれらを言語間リンク辞書において対応する Wikidata エンティティに変換した。

また、エンティティの型を抽出する際は Xiong らの研究 [92] に倣い、Wikidata の “instance of” の項目から抽出した。この際、複数の型が得られる場合はランダムに 1 つサンプリングした。正例エンティティに対して型が同じであり、同じページ上のハイパーリンクエンティティとして登場しないエンティティの中からランダムに 1 つサンプリングしたエンティティをハードネガティブエンティティとして用いた。

多言語の設定では実験に用いる多言語 STS データセットと 後述する MewsC-16 データセットの両方に存在する 18 の言語（アラビア語、カタルーニャ語、チェコ語、ドイツ語、英語、エスペラント、スペイン語、ペルシア語、フランス語、イタリア語、日本語、韓国語、オランダ語、ポーランド語、ポルトガル語、ロシア語、スウェーデン語、トルコ語）の Wikipedia からエンティティ-文データをそれぞれ抽出し、各言語からそれぞれ 5 万データをランダムにサンプリングした。また、単言語の設定では同じ版の英語の Wikipedia からエンティティ-文データを 100 万件サンプリングした。

4.4.3 Semantic Textual Similarity

多言語 STS のデータセットとして STS2017 [59] に含まれる EN-EN、AR-AR、ES-ES、EN-AR、EN-ES、EN-TR ペアを利用した。それらに加えて、Reimers ら [70] が Google 翻訳によって作成した EN-FR、EN-DE、EN-IT、EN-NL ペアも利用した。

また、英語の STS データセットとしてベンチマークとして頻繁に用いられている 7 つの STS データセット（STS 2012–2016 [93, 94, 95, 96, 97]、STSBenchmark [59]、 SICKRelatedness [98]）を利用した。

Gao ら [80] の設定に倣い全ての STS タスクでは文表現から直接コサイン類似度を計測し、相関の計測にはスピアマンの順位相関係数を用いた。さらに STS 各データセット内で全体の STS 値とコサイン類似度から相関係数を計測する “all” の設定を採用した。

4.4.4 Short Text Clustering

提案手法による多言語文表現が大まかなカテゴリ構造情報が埋め込まれているかを確認するため、多言語 Short Text Clustering のデータセットを構築した。

[14] `https://polyglot.readthedocs.io/en/latest/Tokenization.html`

Language	Sentences	Labels
ar	2,243	11
ca	3,310	11
cs	1,534	9
de	6,398	8
en	12,892	13
eo	227	8
es	6,415	11
fa	773	9
fr	10,697	13
ja	1,984	12
ko	344	10
pl	7,247	11
pt	8,921	11
ru	1,406	12
sv	584	7
tr	459	7

表 4.4: MewsC-16 データセットの統計値

このデータセットの構築のために Wikinews からトピックごとに文を収集した。具体的には英語版 Wikinews のテーマ別記事一覧ページ[15]にあるカテゴリ名から複数言語にリンクが存在する 13 件のカテゴリ名（Science and technology, Politics and conflicts, Environment, Sports, Health, Crime and law, Obituaries, Disasters and accidents, Culture and entertainment, Economy and business, Weather, Education, Media）を選択しクラスタのラベルとした。各言語においてこの英語ラベルに対応するラベルに所属するページを収集した。次に各ページのテキストに対して `Wikiextractor`[16] を用いて整形を行った後に `polyglot` 文トークナイザにより文分割を行った。文書中の最初の文が記事の全体の意味をよく表すトピックセンテンスになっていると仮定 [99, 100] し、最初の文を利用した。。最終的に得られたデータセットの統計値を表 4.4 に示す。

また単言語の設定では Zhang らの研究 [79] で利用されている 8 つの short text clustering データセットを用いる。このデータセットはニュース記事やツイートなど幅広いドメインを含むテキストクラスタリングデータセットである。データセットの統計値を表 4.5 に示す。

実験では Zhang らの設定 [79] に倣い、K-Means [83] により文表現に対してクラスタリングを行い、Hungarian アルゴリズム [84] によりクラスタリングの正解率を計算した。異なるシード値で 3 回実験を行い、その平均正解率を示す。

[15] `https://en.wikinews.org/wiki/Category:News_articles_by_section`
[16] `https://github.com/attardi/wikiextractor`

Dataset	Datasize	Label	Category
AgNews (AG)	8K	23	4
StackOverflow (SO)	20K	8	20
Biomedical (Bio)	20K	13	20
SearchSnippets (SS)	12K	18	8
GooglenewsTS (G-TS)	11K	28	152
GooglenewsS (G-S)	11K	22	152
GooglenewsT (G-T)	11K	6	152
Tweet (Tweet)	5K	8	89

表 4.5: Short Text Clustering データセットの統計値

4.4.5　対訳コーパスマッチング

さらに EASE による多言語文表現の汎用性を確かめるために対訳コーパスマッチングタスクでの評価を行った。Tatoeba データセットにおいてクエリとなる英語文（もしくは目的言語の文）の文表現と最もコサイン類似度の高い目的言語（もしくは英語）の文を抽出し、その文がクエリ文の翻訳文と一致しているかの正解率を計測した。

4.4.6　言語間転移タスク

提案手法による多言語文表現を特徴量として用いた言語間転移学習が可能かを調査するため、ゼロショット言語間文書分類タスクである MLDoc を用いた評価を行った。具体的には英語学習データを用いて多言語文表現モデルから得られる文表現を入力の特徴量とする一層の全結合層からなる線形分類器を学習する。その分類器を用いて各言語のテストデータにおける性能を検証する。データとしては MLDoc で提供されている 1000 文の英語データが含まれる `english.train.1000` と `english.dev` をそれぞれ学習データ、検証データとして用い、その他の言語のテストデータで評価を行った。なお、全てのモデルで検証データにてバッチサイズ $\in \{32, 64, 128\}$ と学習率 $\in \{0.1, 0.01, 0.001\}$ のハイパラーパラメータ調整を行っている。

4.4.7　Wikipedia2vec

提案手法では Wikidata エンティティ表現は事前に学習したエンティティ表現で初期化した。具体的にはオープンソース Wikipedia2Vec [23] を利用し、2019 年 1 月版の英語 Wikipedia からベクトルの次元をベースモデルの隠れ表現と同じ 768 と設定し、その他はデフォルトの設定で学習した。

Pooler	SimCSE	EASE
[CLS]		
w/ MLP	63.0	65.0
w/ MLP (train)	72.0	73.3
w/o MLP	72.0	73.4
mean pooling	**72.1**	**73.8**

表 4.6: 多言語設定の SimCSE、EASE における異なるプーリング手法の比較。結果は STS-B と SICK-R の検証データにおけるスピアマンの相関係数の平均を表す。

Model	Batch size	Learning Rate	τ	λ
SimCSE-mBERT$_{\text{base}}$	128	3e-05	-	-
SimCSE-XLM-R$_{\text{base}}$	128	3e-05	-	-
EASE-BERT$_{\text{base}}$	64	3e-05	100	0.01
EASE-RoBERTa$_{\text{base}}$	128	5e-05	100	0.01
EASE-mBERT$_{\text{base}}$	256	5e-05	10	0.01
EASE-XLM-R$_{\text{base}}$	64	3e-05	10	0.01

表 4.7: ハイパーパラメータの値。

4.4.8　詳細設定

EASE は事前学習済み汎用言語モデルがライブラリとして提供されている `transformers`[17]の実装を拡張する形式で PyTorch により実装した。

文表現を EASE もしくは SimCSE から獲得する際のプーリングの手法は、多言語の設定では最終層の隠れ表現の各トークンの平均を利用する mean pooling、単言語の設定では Gao ら [80] に倣い [CLS] トークンに対応する最終層の隠れ表現を用いた。なお、多言語の設定におけるプーリングの手法は、検証データにおいて他のいくつかの 手法と比較し最も良い結果であった mean pooling を採用している (表 4.6)。

多言語設定においては検証データとして STS-B、SICK-R を用い、それらのスピアマンの順位相関係数の値の平均値を基準にハイパーパラメータを調整した。また、単言語の設定では検証データとして STS-B を用いた。250 学習ステップごとに検証データにおける各評価指標の値を計算し、最も値が高いチェックポイントのモデルを最終的にテストデータで推論を行うモデルとした。

この際、検証データにおいて EASE モデルでは温度付き Softmax 関数における温度 $\tau \in \{100, 10, 1\}$ とエンティティを用いた Contrastive learning と自己教師あり学習のバランスを定める $\lambda \in \{0.01, 0.01, 1\}$ のグリッドサーチを行っている。また全てのモデルにてバッチサイズ $\in \{64, 128, 256, 512\}$ と学習率 $\in \{3e-05, 5e-05\}$ のグリッドサーチを行っている。実際に選択されたハイパーパラメータの値を表 4.7 に示す。

[17] `https://huggingface.co/docs/transformers/index`

Model	EN-EN	AR-AR	ES-ES	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
$\text{mBERT}_{\text{base}}$(avg.)	54.4	50.9	56.7	18.7	33.9	16.0	21.5	33.0	34.0	35.3	35.4
$\text{SimCSE-mBERT}_{\text{base}}$	77.7	**63.7**	77.4	25.8	53.9	**28.7**	43.8	48.9	52.2	50.4	52.3
$\text{EASE-mBERT}_{\text{base}}$	**79.3**	62.8	**79.4**	**31.6**	**59.8**	26.4	**53.7**	**59.2**	**59.4**	**60.7**	**57.2**
$\text{XLM-R}_{\text{base}}$(avg.)	52.2	25.5	49.6	15.7	21.3	12.1	10.6	16.6	22.9	23.9	25.0
$\text{SimCSE-XLM-R}_{\text{base}}$	79.6	64.2	**80.8**	29.3	58.8	**38.9**	44.3	**55.2**	49.0	59.1	55.9
$\text{EASE-XLM-R}_{\text{base}}$	**80.6**	**65.3**	80.4	**34.2**	**59.1**	37.6	**46.5**	51.2	**56.6**	**59.5**	**57.1**

表 4.8: STS2017 の STS 値と文表現のコサイン類似度のスピアマンの順位相関係数

Model	ar	ca	cs	de	en	eo	es	fa	fr	ja	ko	pl	pt	ru	sv	tr	Avg.
$\text{mBERT}_{\text{base}}$ (avg.)	27.0	27.2	**44.3**	36.2	37.9	25.6	**41.1**	35.0	25.9	44.2	31.0	35.0	30.1	23.4	28.9	34.9	33.0
$\text{SimCSE-mBERT}_{\text{base}}$	30.1	26.9	41.3	32.5	37.3	27.2	36.2	**36.9**	29.0	48.9	33.9	37.6	37.9	**27.1**	26.9	35.3	34.1
$\text{EASE-mBERT}_{\text{base}}$	**31.9**	**29.6**	38.8	**38.5**	**30.2**	**34.5**	37.2	36.7	**30.4**	**49.3**	**36.2**	**40.0**	**41.0**	27.0	**30.5**	**44.7**	**36.0**
$\text{XLM-R}_{\text{base}}$ (avg.)	**26.0**	24.7	28.2	29.4	23.0	23.5	22.1	36.6	23.6	38.8	22.0	24.2	32.8	18.0	**33.2**	26.0	27.0
$\text{SimCSE-XLM-R}_{\text{base}}$	24.6	26.3	34.6	28.6	33.4	31.7	32.9	35.9	29.1	41.1	31.1	33.1	30.0	26.0	32.9	37.2	31.8
$\text{EASE-XLM-R}_{\text{base}}$	25.3	**26.7**	**43.2**	**37.0**	**34.9**	**34.2**	**37.2**	**42.4**	**32.0**	**46.0**	**32.8**	**41.6**	**33.4**	**31.3**	27.2	**41.8**	**35.4**

表 4.9: MewsC-16 における多言語クラスタリングの各言語の正解率

4.5　実験結果

4.5.1　多言語設定の結果

まず、多言語設定における実験の結果について述べる。

表 4.8 に多言語 STS の結果を示す。実験結果から EASE は、全ての言語対にてそれらのベースモデルである mBERT、XLM-R を大幅に超える平均性能を発揮し、mBERT では 35.4 から 57.2、XLM-R では 25.0 から 57.1 の向上を確認した。さらに教師なし手法の代表的な手法である SimCSE よりも EASE は平均的に高い性能を発揮することが確認された。

表 4.9 に多言語 STC の結果を示す。多言語 STC においても EASE はベースモデルである mBERT、XLM-R を超える性能を発揮し、特に EASE-XLM-R は XLM-R から 8.4% の正解率の向上を達成した。さらに EASE は多言語 STS と同様に SimCSE よりも平均的に性能が高いことが確認された。これらの結果は EASE による文表現が多言語の細かい意味的含意や矛盾を推論できると同時に、より大まかなカテゴリ構造情報を捉えていることを示している。

表 4.10 に対訳コーパスマッチングの結果を示す。このタスクにおいては EASE はそれらのベースモデルや SimCSE と比較し、大幅な性能の向上が確認された。特に $en \rightarrow xx$ では EASE の平均正解率が SimCSE と比較しそれぞれ mBERT で 37.7% から 54.2%、XLM-R で 62.9% から 67.5% まで向上しており、他の多言語タスクと比較するとこのタスクでは特に目覚ましい性能の向上を達成している。対訳コーパスマッチングは STS や STC、言語間転移タスクと比較すると、より直接的に同じ意味を持つ異なる言語の文が文表現上で近傍に存在するかを計測するタスクであり、この実験の結果は EASE の学習がそのような言語間で意味的に類似した文を紐付けることに特に秀でていることを示唆している。

表 4.11 に MLDoc における言語間文書分類の結果を示す。このタスクにおいては平均的には EASE はそれらのベースモデルを超える性能を発揮しており、特に EASE-mBERT は mBERT と比較すると 63.7% から 69.5% まで正解率が向上している。また SimCSE モデルに注目すると

Model	ar	ca	cs	de	eo	es	fr	it	ja	ko	nl	pl	pt	ru	sv	tr	Avg.
	en → xx																
$\text{mBERT}_{\text{base}}$(avg.)	17.8	44.6	29.6	59.6	12.0	52.4	50.8	45.7	38.0	33.8	53.2	37.5	54.7	49.1	39.7	27.8	40.4
$\text{SimCSE-mBERT}_{\text{base}}$	14.8	42.5	25.6	52.8	18.7	49.3	48.8	46.1	41.2	27.9	46.4	31.5	47.2	46.1	39.9	24.9	37.7
$\text{EASE-mBERT}_{\text{base}}$	**29.3**	**63.0**	**43.0**	**72.4**	**25.7**	**65.4**	**65.9**	**61.6**	**58.9**	**45.2**	**66.8**	**52.5**	**65.0**	**61.8**	**55.7**	**35.2**	**54.2**
$\text{XLM-R}_{\text{base}}$(avg.)	9.1	9.1	11.4	41.8	3.5	35.5	26.3	19.9	13.7	16.7	42.1	18.2	34.5	36.2	39.3	11.5	23.1
$\text{SimCSE-XLM-R}_{\text{base}}$	35.1	58.5	54.0	77.4	46.6	70.0	69.3	67.3	56.8	53.9	70.6	64.3	77.4	71.8	73.2	60.9	62.9
$\text{EASE-XLM-R}_{\text{base}}$	**39.3**	**63.0**	**59.0**	**85.5**	**53.7**	**74.3**	**73.4**	**69.0**	**67.0**	**58.8**	**76.4**	**68.6**	**79.8**	**75.1**	**77.0**	**60.7**	**67.5**
	xx → en																
$\text{mBERT}_{\text{base}}$(avg.)	23.4	53.8	36.0	65.9	12.3	62.9	60.3	55.8	39.1	32.4	56.3	42.8	62.3	53.7	51.8	32.4	46.3
$\text{SimCSE-mBERT}_{\text{base}}$	19.5	55.0	35.3	61.6	20.3	60.6	57.1	54.3	42.0	33.7	57.8	42.1	59.1	51.8	49.6	29.4	45.6
$\text{EASE-mBERT}_{\text{base}}$	**34.9**	**70.0**	**52.4**	**75.9**	**26.5**	**74.7**	**67.5**	**68.9**	**59.4**	**48.4**	**71.5**	**58.3**	**73.1**	**67.0**	**63.0**	**41.0**	**59.5**
$\text{XLM-R}_{\text{base}}$(avg.)	11.4	21.4	21.5	57.4	11.5	37.2	35.3	31.3	16.3	21.9	48.3	29.9	49.4	38.6	46.2	24.2	31.4
$\text{SimCSE-XLM-R}_{\text{base}}$	39.7	65.5	61.1	82.3	51.9	74.4	72.3	71.4	59.9	54.3	75.6	70.0	79.7	73.8	76.2	**66.1**	67.1
$\text{EASE-XLM-R}_{\text{base}}$	**45.8**	**67.2**	**68.6**	**88.8**	**58.4**	**77.5**	**74.8**	**72.6**	**69.4**	**62.2**	**79.4**	**75.2**	**81.3**	**77.8**	**81.4**	61.1	**71.3**

表 4.10: Tatoeba データセットにおける対訳コーパスマッチングの正解率

Model	en (dev)	de	es	fr	it	ja	ru	zh	Avg.
$\text{mBERT}_{\text{base}}$ (avg.)	**89.5**	68.0	68.1	70.6	62.7	61.2	61.5	69.6	65.9
$\text{SimCSE-mBERT}_{\text{base}}$	88.4	62.3	**73.2**	78.2	64.3	**63.7**	61.3	**75.0**	68.3
$\text{EASE-mBERT}_{\text{base}}$	89.0	**69.9**	69.2	**80.1**	**66.8**	62.8	**64.4**	73.2	**69.5**
$\text{XLM-R}_{\text{base}}$ (avg.)	**90.9**	**82.7**	**79.8**	72.1	72.5	71.1	69.6	71.4	74.2
$\text{SimCSE-XLM-R}_{\text{base}}$	90.7	74.9	74.1	81.5	70.3	71.7	70.1	76.6	74.2
$\text{EASE-XLM-R}_{\text{base}}$	90.6	77.9	75.6	**83.9**	**72.6**	**72.8**	**71.1**	**81.6**	**76.5**

表 4.11: MLDoc データセットにおける言語間文書分類タスクの正解率

SimCSE-mBERT は EASE-mBERT にわずかに勝る結果が観測されたが、SimCSE-XLM-R はベースモデルの XLM-R とおおよそ変わらない性能になっており、一貫した性能の向上は確認されなかった。

以上の多言語タスクにおけるいくつかの結果は EASE におけるエンティティの言語非依存性を活用した多言語文表現の学習が有効であることを示唆している。

4.5.2　単言語設定の結果

次に、単言語設定における実験の結果について述べる。

表 4.12 に英語 STS の結果を示す。全体としては既存の教師なし学習による文表現学習手法と比較すると EASE が最も高い相関係数を達成している。しかし、既存の state of the art の手法である SimCSE と比較すると平均性能は BERT で 0.6 ポイント、RoBERTa で 0.3 ポイントとわずかな改善に留まっており、タスク別に見ても一貫して SimCSE より EASE の性能が優れているわけではなかった。本手法におけるエンティティによる Contrastive learning はエンティティを軸に関連する様々な文をエンティティ表現上に埋め込む手法であるため、似た意味を持つ文の文表現を近づける効果は期待されるが、文の細かい意味的矛盾を推論する効果は期待されない。さらに多言語設定のように文表現を言語非依存にすることもないため、STS では特に大きな改善が観測されなかったと考えられる。

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)	55.1	70.7	59.7	68.3	63.7	58.0	53.8	61.3
$\text{BERT}_{\text{base}}$ (avg.)	30.9	59.9	47.7	60.3	63.7	47.3	58.2	52.6
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.7	59.4	49.7	66.0	66.2	53.9	62.1	56.7
$\text{BERT}_{\text{base}}$-flow	58.4	67.1	60.9	75.2	71.2	68.7	64.5	66.6
$\text{BERT}_{\text{base}}$-whitening	57.8	66.9	60.9	75.1	71.3	68.2	63.7	66.3
IS-$\text{BERT}_{\text{base}}^{\heartsuit}$	56.8	69.2	61.2	75.2	70.2	69.2	64.3	66.6
CT-$\text{BERT}_{\text{base}}$	61.6	76.8	68.5	77.5	76.5	74.3	69.2	72.1
SimCSE-$\text{BERT}_{\text{base}}$	68.4	**82.4**	**74.4**	80.9	78.6	76.9	**72.2**	76.3
EASE-$\text{BERT}_{\text{base}}$	**72.8**	81.8	73.7	**82.3**	**79.5**	**78.9**	69.7	**76.9**
$\text{RoBERTa}_{\text{base}}$ (avg.)	32.1	56.3	45.2	61.3	62.0	55.4	62.0	53.5
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.9	58.7	49.1	65.6	61.5	58.6	61.6	56.6
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	52.4	75.2	65.5	77.1	78.6	72.4	**68.6**	70.0
SimCSE-$\text{RoBERTa}_{\text{base}}$	68.7	**82.6**	**73.6**	81.5	**80.8**	**80.5**	67.9	76.5
EASE-$\text{RoBERTa}_{\text{base}}$	**70.9**	81.5	73.5	**82.6**	80.5	80.0	68.4	**76.8**

表 4.12: 英語 STS データセットにおけるの STS 値と文表現のコサイン類似度のスピアマンの順位相関係数

Model	AG	Bio	G-S	G-T	G-TS	SO	SS	Tweet	Avg.
GloVe embeddings (avg.)	83.2	30.7	59.0	58.3	67.4	29.9	70.4	52.1	56.4
$\text{BERT}_{\text{base}}$ (avg.)	79.8	32.5	55.0	47.0	62.4	21.7	64.0	44.6	50.9
CT-$\text{BERT}_{\text{base}}$	79.2	**38.7**	**65.5**	**60.7**	**69.8**	67.9	55.5	**55.2**	61.6
SimCSE-$\text{BERT}_{\text{base}}$	74.4	34.3	59.5	57.8	64.4	49.6	64.3	52.1	57.1
EASE-BERT$_{\text{base}}$	**85.8**	36.2	60.5	60.4	67.0	**68.1**	**71.7**	54.8	**63.1**
$\text{RoBERTa}_{\text{base}}$ (avg.)	66.5	26.6	47.9	42.8	58.3	16.7	30.0	38.6	40.9
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	**80.7**	**41.0**	**65.2**	**60.5**	**69.6**	32.9	**73.6**	**56.8**	**60.0**
SimCSE-$\text{RoBERTa}_{\text{base}}$	69.8	37.3	60.0	58.0	66.6	69.3	48.3	50.0	57.4
EASE-RoBERTa$_{\text{base}}$	69.4	39.3	60.7	57.7	66.3	**73.9**	49.4	51.8	58.6

表 4.13: 英語 STC データセットにおけるクラスタリングの正解率

表 4.13 に英語 STC の結果を示す。全体として、EASE-BERT は既存の教師なし学習による文表現学習手法と比較すると最も高い正解率を達成しており、既存の最も正解率の高い CT-BERT と比較しても 2.0% の向上が確認された。また state-of-the-art モデルである SimCSE と比較すると、BERT では 57.1% から 63.1%、RoBERTa では 57.4% から 58.6% と明確な性能の改善が観測された。この結果は EASE が大まかなカテゴリ構造情報を捉えた文表現の学習に有効であることを示唆している。

以上の英語 STS、STC の結果から EASE は state-of-the-art モデルである SimCSE と比較すると文の細かい意味的含意・矛盾を識別する能力を維持しつつ、より大まかなカテゴリ構造情報を捉えることが可能な文表現学習だと言える。

Setting	EASE-BERT$_{\text{base}}$ STS avg.	EASE-RoBERTa$_{\text{base}}$ STS avg.	EASE-mBERT$_{\text{base}}$ mSTS avg.	EASE-XLM-R$_{\text{base}}$ mSTS avg.
Full model	**76.9**	**76.8**	**57.2**	**57.1**
w/o self-supervised	65.3	66.1	49.3	53.1
w/o hard negative	75.3	76.1	53.8	52.7
w/o Wikipedia2Vec	73.8	76.3	52.1	54.3
w/o all (vanilla model)	31.4	43.6	35.4	25.0

表 4.14: Ablation study

4.6 分析

4.6.1 Ablation study

本項では EASE モデルを構成する各要素の性能へどの程度寄与しているかを調査した。具体的には今回の実験で学習した 4 つのモデル（EASE-BERT$_{\text{base}}$、EASE-RoBERTa$_{\text{base}}$、EASE-mBERT$_{\text{base}}$、EASE-XLM-R$_{\text{base}}$）に対して、(1) 自己教師あり Contrastive learning の目的関数を利用せず学習したモデル（w/o self-supervised）(2) hard negative を利用せず学習したモデル（w/o hard negative)、(3) Wikipedia2Vec による初期化を行わず学習したモデル（w/o Wikipedia2Vec)、(4) 全ての要素を取り除いたモデル（vanilla model）[18]の 4 つを 4.4.3 項と同様に STS の評価値で比較した。

結果を表 4.14 に示す。まず全ての要素を用いて学習した Full model と比較すると、いずれの要素を取り除いても性能が低下していることがわかり、全ての要素が性能に寄与していることが確認された。興味深い点としてエンティティの Contrastive learning のみを行った設定（w/o self-supervised）においてもモデルは vanilla model より大幅に性能が向上している。特に多言語設定における性能への寄与は著しく、XLM-R に関しては 25% から 53.1% まで向上している。この結果は複数の異なる言語の似た意味を持つ文を共通のエンティティベクトルに近づける埋め込むエンティティの Contrastive learning が多言語文表現学習に有効であることを裏付けている。

4.6.2 Uniformity と Alignment

提案手法である EASE が何故ベースモデルの性能を改善させたかについて調査するため、alignment と uniformity を計測した。具体的には単言語設定における BERT とそれに基づく EASE、SimCSE モデルの各モデルにて STS-B の検証データにおける alignment と uniformity を算出した。さらに多言語設定においても 4.4.3 節で扱った多言語 STS データセットを用いて mBERT とそれに基づく SimCSE、EASE における alignment と uniformity を計測した。なお、このデータセットにおいては各言語ペアの alignment と uniformity の平均スコアを算出している。

図 4.2 にその結果を示す。SimCSE と EASE はベースモデルと比較すると全ての言語ペアで uniformity の大幅な改善に成功していることがわかる。すなわちどちらの学習もベースラインと比

[18] このモデルは事前学習済み汎用言語モデルに対して一切ファインチューニングを行っていない状態で提案手法と同じ pooler により文表現を獲得するモデルである。

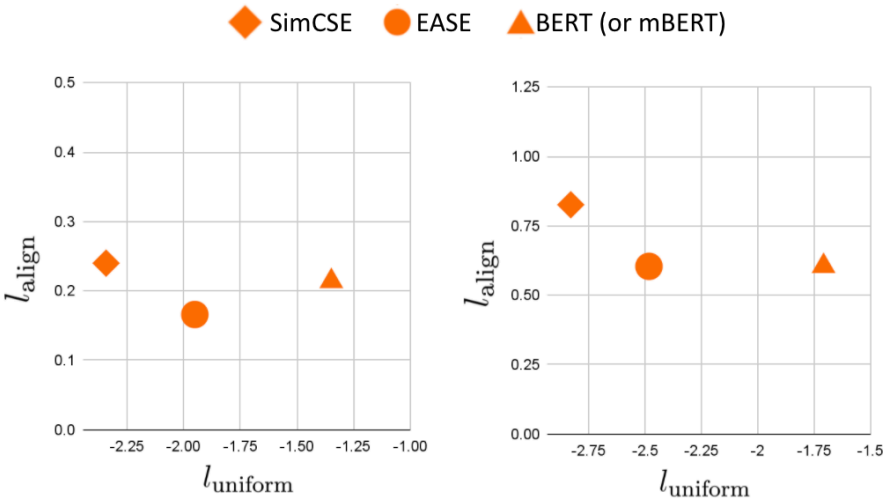


図 4.2: alignment と uniformity のプロット図。左側が単言語設定、右側が多言語設定を表す。

較し文表現をより均等に分布させることに成功している。次に SimCSE と EASE を比較すると、EASE は alignment においては SimCSE より良い数値だが、uniformity においては劣る数値であることが観測された。この結果はエンティティを用いた Contrastive learning は文表現を均等に分布させる効果はあまりないが、意味的に類似した意味を持つ文の文表現を近づける効果はあり、これにより文表現が改善された可能性を示唆している。

4.6.3　教師ありモデルへの EASE の適用

大規模な対訳コーパスで学習された既存の多言語文表現モデルは学習に用いた資源には含まれていない言語に対しては性能を発揮できない場合がある。例えば LaBSE モデルは約 60 億対の対訳コーパスを用いてモデルの学習を行うが、対訳コーパスに含まれていないカビル語に対しては Tatoeba における対訳コーパスマッチングで 5% 未満の性能しか発揮できない。

一方で EASE では 300 以上の比較的多くの言語で豊富に存在する Wikipedia テキストコーパスを活用するため、対訳コーパスや XNLI など特定のいくつかの 言語でしか構築されていない資源よりも広い言語にて適用が可能である。したがって EASE による多言文表現学習は対訳コーパスによる学習ではカバーされていない低資源言語における性能を補う効果が期待される。

この点を検証するため、LaBSE に対してさらに EASE のフレームワークで fine-tune する実験を行った。具体的には LaBSE の学習対訳コーパスに含まれていない言語の中で性能が特に低く、かつ Wikipedia のデータが存在する 5 つの言語（カビル語、パンパンガ語、 コーンウォール語、ブルトン語、マリ語）において LaBSE と英語+該当低資源言のデータをそれぞれ 5,000 ペアずつ用いて EASE による fine-tune を LaBSE に対して行ったモデル（LaBSE + EASE (en + xx)）の Tatoeba における性能を比較した。

この実験の結果を図 4.3 に示す。全ての言語群において、LaBSE を EASE により追加学習を行

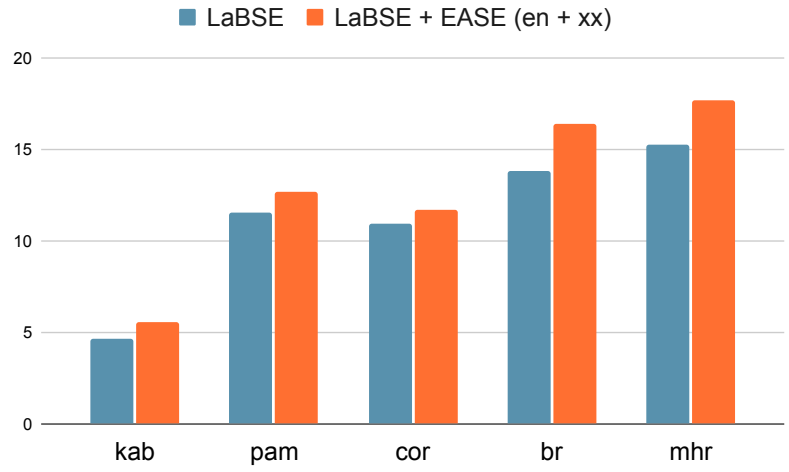


図 4.3: LaBSE モデルを EASE によって fine-tune した場合の Tatoeba の対訳コーパスマッチングにおける両方向の正解率の平均。

うことで性能の改善に成功している。これは LaBSE が約 60 億対の対訳コーパスで既に学習されていることを考慮すると驚くべき結果である。この結果は特に対訳コーパスが少ない低資源言語において、EASE による Wikipedia を活用した学習を対訳コーパスを用いた多言語文表現モデルと組み合わせることで高性能な多言語文表現モデルを学習できる可能性を示唆している。

4.7　結論

4.7.1　まとめ

本研究では Wikipedia 知識ベースを活用し、文をそれと関連するエンティティを近づけ、関連しないエンティティとは遠ざける Contrastive learning による文表現学習手法、EASE を提案した。この手法では言語に非依存なエンティティを軸に関連する様々な言語の文を共通のベクトルの近傍に埋め込まれるため、特に多言語文表現の学習に有効である。実験では単言語設定、多言語設定でそれぞれ EASE による文表現を学習し、特に多言語の自然言語処理タスクで既存の教師なし文表現より優れた性能を発揮することが確認された。また、EASE の活用事例として、低資源言語において既存の教師ありモデル LaBSE の性能を補えることを示し今後の多言語文表現モデルにエンティティ情報を組み込んでいくことの有効性を示唆した。

4.7.2　今後の課題

本研究では主に対訳コーパスを用いない設定での、EASE の多言語文表現学習の有用性を示したが、既存の対訳コーパスなどの言語資源を利用した多言語文表現モデルと比較すると EASE の性能

は劣る。また、4.6.3 節では LaBSE に対して EASE を適用させる実験を行ったが、state of the art の性能を目指すことを前提として様々な言語資源を活用した最適なモデル構築を行っているわけではない。そこで実用性を考慮した高性能な多言語文表現モデルを構築するために対訳コーパスや XNLI などの言語資源と、本研究による Wikpedia を活用したエンティティ Contrastive learning の学習を組み合わせるような手法の探究が今後の展望として考えられる。

また、本研究におけるエンティティの Contrastive learning は、Wikipedia のハイパーリンク情報から得られるエンティティ-文データセットのような、文と文中のエンティティがアノテーションされた言語資源でしか学習ができないという制限がある。そこで EASE をより幅広いドメインで活用するため、エンティティリンキングシステムにより文からエンティティを抽出することで、エンティティ-文データセットを生成し EASE の学習を行うことが可能か検証することが今後の展望として考えられる。

第5章　おわりに

本論文では、知識ベースを活用した多言語モデルによるゼロショット言語間文書分類タスクと多言語文表現学習の性能改善を試みた。ゼロショット言語間文書分類における研究では、文書中から言語に非依存であり文書のトピックを良く捉えるエンティティを抽出し、それらを分類タスクの入力特徴量として利用する多言語エンティティモデルを提案した。このモデルは複数の言語間文書分類タスクにおける多くの言語にて既存のモデルと比較し性能の向上を達成した。多言語文表現学習における研究では、文とその文に関連するエンティティに近づけ、関連しないエンティティとは遠ざける EASE を提案した。実験では様々な言語処理タスクにて EASE による文表現を評価し、既存のモデルと比べ言語に依存せず文の意味を上手く捉えていることが確認された。以上の 2 つの研究は知識ベースが多言語モデルの学習資源として有用であることを示唆している。

本研究で提案したエンティティとテキスト情報のみを用いた多言語モデルの性能は対訳コーパスや XNLI などの言語資源を大量に用いて学習した多言語モデルには様々なタスクの性能で劣り、実用に耐えうる性能であるとは言い難い。とはいえ、エンティティの（1）文書のトピックを捉える点、（2）低資源言語でも手に入る点は対訳コーパスと差別化できる特徴である。従って今後の展望としては、エンティティや対訳コーパスを始めとする様々な種類の言語資源を活用し、それぞれの強みを活かしたより汎用的な多言語モデルの構築が考えられる。

また、本研究で提案した 2 つの多言語モデルは、Wikipedia にデータが存在する言語のみにしか適用できないという制限がある。しかし、Wikipedia は 2022 年 1 月現在 300 以上の言語で構成され、その半数以上が 1 万記事以上から成り、比較的広い言語で豊富に存在する資源である。[19] さらに、知識ベースは世界中の編集者によって現在も拡張が進められており、低資源言語の充実化に加えて、新たな言語を追加していく取り組みも行われている。[20] 従って知識ベースは将来的により幅広い言語で充実した言語資源になることが考えられ、それに伴い、将来的に本研究における提案モデルはより幅広い言語に対して、より高い性能を発揮することが期待される。

[19] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[20] https://incubator.wikimedia.org/wiki/Incubator:Main_Page

参考文献

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, 2017.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[5] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, June 2018.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

[8] 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, June 2019.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, November 2018.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 235–243, August 2009.

[13] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 572–581, July 2012.

[14] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[15] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[16] Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, June 2015.

[17] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, July 2018.

[18] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, Vol. 65, No. 1, p. 569–630, may 2019.

[19] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, July 2019.

[20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, July 2020.

[21] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, July 2020.

[22] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, October-November 2018.

[23] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 23–30, October 2020.

[24] Haotian Chen, Xi Li, Andrej Zukov Gregoric, and Sahil Wadhwa. Contextualized end-to-end neural entity linking. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 637–642, Suzhou, China, December 2020. Association for Computational Linguistics.

[25] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 58–68, August 2017.

[26] Ikuya Yamada and Hiroyuki Shindo. Neural attentive bag-of-entities model for text classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pp. 563–573, November 2019.

[27] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, July 2020.

[28] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4483–4499, November 2020.

[29] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, April 2017.

[30] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, p. 1301–1306, 2006.

[31] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, p. 830–835, 2008.

[32] Sapna Negi and Michael Rosner. UoM: Using explicit semantic analysis for classifying sentiments. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 535–538, June 2013.

[33] Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. Cross-lingual dataless classification for many languages. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, p. 2901–2907, 2016.

[34] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 43–54, November 2019.

[35] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, May 2018.

[36] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 58–68, June 2014.

[37] Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. Overview of shinra2020-ml task. In *Proceedings of the NTCIR-15 Conference*, 2020.

[38] Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. Bridging the domain gap in cross-lingual document classification. *arXiv preprint arXiv:1909.07009*, 2019.

[39] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[40] Phillip Keung, Yichao Lu, and Vikas Bhardwaj. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1355–1360, November 2019.

[41] Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 1541–1544, 2020.

[42] Xin Dong and Gerard de Melo. A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6306–6310, November 2019.

[43] Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5702–5707, November 2019.

[44] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, Vol. 32, pp. 7059–7069, 2019.

[45] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1441–1451, July 2019.

[46] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6442–6454, November 2020.

[47] Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4937–4951, November 2020.

[48] Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, and Georg Rehm. Enriching BERT with Knowledge Graph Embedding for Document Classification. In *Proceedings of the GermEval 2019 Workshop*, Erlangen, Germany, 2019.

[49] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 803–818, November 2020.

[50] Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1952–1957, August 2021.

[51] Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1020–1030, June 2013.

[52] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pp. 233–242, 2007.

[53] Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 549–554, November 2020.

[54] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 597–610, 2019.

[55] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[56] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 20–22 Jun 2016.

[57] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 176–194, 03 2021.

[58] Zellig Harris. Distributional structure. *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

[59] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, August 2017.

[60] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[61] Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9099–9113, November 2021.

[62] Sosuke Nishikawa, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. A multilingual bag-of-entities model for zero-shot cross-lingual text classification. *arXiv preprint arXiv:2110.07792*, 2021.

[63] Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. mluke: The power of entity representations in multilingual pretrained language models. *arXiv preprint arXiv:2110.08151*, 2021.

[64] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 28, 2015.

[65] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[66] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, September 2017.

[67] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, November 2019.

[68] Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, May 2018.

[69] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 597–610, March 2019.

[70] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, November 2020.

[71] Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 5370–5378, 2019.

[72] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Learning distributed representations of texts and entities from knowledge base. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 397–411, 2017.

[73] Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. DBpedia abstracts: A large-scale, open, multilingual NLP training corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3339–3343, May 2016.

[74] Yury Zemlyanskiy, Sudeep Gandhe, Ruining He, Bhargav Kanagal, Anirudh Ravula, Juraj Gottweis, Fei Sha, and Ilya Eckstein. DOCENT: Learning self-supervised entity representations from large document collections. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2540–2549, April 2021.

[75] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2, pp. 1735–1742, 2006.

[76] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

[77] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, p. 767–776, 2017.

[78] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5065–5075, August 2021.

[79] Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5786–5798, November 2021.

[80] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021.

[81] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1442–1459, November 2021.

[82] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 56, pp. 1929–1958, 2014.

[83] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297. University of California Press, 1967.

[84] James R. Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, Vol. 5, No. 1, pp. 32–38, March 1957.

[85] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939, 2020.

[86] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 13–18 Jul 2020.

[87] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[88] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[89] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, October 2014.

[90] Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[91] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 879–895, August 2021.

[92] Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[93] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 385–393, 7-8 June 2012.

[94] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 32–43, June 2013.

[95] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 81–91, August 2014.

[96] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 252–263, June 2015.

[97] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511, June 2016.

[98] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, May 2014.

[99] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, Vol. 16, No. 2, p. 264–285, apr 1969.

[100] P. B. Baxendale. Machine-made index for technical literature: An experiment. *IBM J. Res. Dev.*, Vol. 2, No. 4, p. 354–361, oct 1958.

発表文献

査読付き会議論文

1. A Multilingual Bag-of-Entities Model for Zero-Shot Cross-Lingual Text Classification. Sosuke Nishikawa, Ikuya Yamada, Yoshimasa Tsuruoka and Isao Echizen. The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop (ACL-IJCNLP SRW 2021). https://arxiv.org/abs/2110.07792

査読なし会議

1. エンティティの言語間リンクに基づく多言語モデルの構築. 西川 荘介, 山田 育矢, 鶴岡 慶雅, 越前 功. 言語処理学会 第 27 回年次大会 (2021). https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/P8-14.pdf

謝辞

本研究を進めるにあたって、大変多くの方にお世話になりました。

指導教員である越前先生には計算環境の提供や研究発表の改善など様々な方面で私の研究生活をサポートしていただきました。感謝を申し上げます。

越前研究室の修士学生である、早稲田くん、吉田くんとは、他愛のないことから研究内容に関することまで様々なお話をさせていただき、非常に刺激的な研究生活を送ることができました。

鶴岡研究室の鶴岡先生、李凌寒さんには研究に関して深い議論をしていただいたことで、どのように研究内容を深めていくかが明確になりました。ありがとうございます。

Studio Ousia の山田さんには論文の書き方からモデルの実装の細かい部分まで様々なアドバイスをいただきり、非常に勉強になりました。

お茶の水女子大学の田上さんには日頃の悩みから研究に関する疑問点まで多岐に渡り相談させていただきました。ありがとうございます。

最後に、私の考えを尊重し大学院の学費をサポートしていただいた家族に感謝申し上げます。