Vol.2020-NL-244 No.6
2020/7/3

機械翻訳を用いた 自然言語推論データセットの多言語化

吉越 卓見[1,a)]　河原 大輔[1,†1,b)]　黒橋 禎夫[1,c)]

概要：言語を理解するには、字義通りの意味を捉えるだけでなく、それが含意する意味を推論することが不可欠である。このような推論能力を計算機に与えるために、自然言語推論（NLI）の研究が盛んに行われている。 NLI は、前提が与えられたときに、仮説が成立する（含意）、成立しない（矛盾）、判別できない（中立）かを判断するタスクある。自然言語推論を計算機で解くには数十万規模の前提・仮説ペアのデータセットが必要となるが、これまでに構築された自然言語推論データセットは言語間でその規模に大きな隔たりがある。この状況は、自然言語推論の研究の進展を妨げる要因となっている。このような背景から、本研究では、機械翻訳に基づく、安価かつ高速な自然言語推論データセットの構築手法を提案する。提案する構築手法は二つのステップからなる。まず、既存の大規模な自然言語推論データセットを機械翻訳によって目的の言語に変換する。次に、翻訳によって生じるノイズを軽減するため、フィルタリングを行う。フィルタリングの手法として、評価データと学習データに対し、それぞれ別のアプローチをとる。評価データは、正確さが重要となるため、クラウドソーシングを用い、人手で検証する。学習データは、大規模な自然言語推論データセットでは数十万ペアの問題が存在するため、翻訳文の検証を自動的に行い、効率的にデータをフィルタリングする。本研究では、機械翻訳を用いた逆翻訳による手法と、言語モデルによる手法の二つを提案する。本研究では、SNLI を翻訳対象とし、日本語を対象言語として実験を行った。その結果、評価データが 3,917 ペア、学習データが 53 万ペアのデータセットを構築した。このデータセットは BERT に基づく自然言語推論モデルによって 93.0%の精度で解くことが可能である。

キーワード：自然言語推論, 機械翻訳, クラウドソーシング, フィルタリング

Multilingualization of a Natural Language Inference Dataset Using Machine Translation

Takumi Yoshikoshi[1,a)]　Daisuke Kawahara[1,†1,b)]　Sadao Kurohashi[1,c)]

Abstract: To understand natural language text, it is essential not only to capture the literal meaning but also to infer the meaning it implies. In order to give such inference ability to computers, research on natural language inference (NLI) has been actively conducted. NLI is a task of determining whether a hypothesis holds (entailment), does not hold (contradiction), or cannot be determined (neutral) when a hypothesis is given. In this study, we propose a low-cost and rapid construction method of NLI datasets based on machine translation. The proposed construction method consists of two steps. First, an NLI dataset is translated into a target language by machine translation. Next, filtering is performed to reduce noise caused by translation. As filtering methods, we propose different approaches for evaluation and training data. Quality is important for evaluation data to accurately measure accuracy based on it. Therefore, filtering is performed manually using crowdsourcing. This manual filtering judges that a translated sentence is a natural sentence in the target language, and that the original NLI relation is correct. Since training data consists of hundreds of thousands of problems in a large NLI dataset, we propose two methods: a back translation-based method using machine translation and a language model-based method. In this study, we conducted an experiment with SNLI as the translation target and Japanese as the target language. As a result, we succeeded in constructing an NLI dataset with 3,917 pairs of evaluation data and 550,000 pairs of training data. A BERT-based NLI model traind on the obtained dataset achieved an accuracy of 93.0%.

Keywords: Natural Language Inference, Machine Translation, Crowdsourcing, Filtering

1. はじめに

言語を理解するには、字義通りの意味を捉えるだけでなく、それが含意する意味を推論することが不可欠である。こうした推論能力を計算機に与えることを目的として、自然言語推論 (Natural Language Inference) の研究が盛んに行われている。自然言語推論とは、前提とそれに対する仮説が与えられたときに、その仮説が成り立つ (含意) か、成り立たない (矛盾) か、判断できない (中立) かを判定するタスクである。例 (1) の問題をみると、前提に対して仮説が正しく、含意となる。

(1) a. 前提：複数の男性がサッカーをしている。
b. 仮説：スポーツをしている人がいる。
c. ラベル：含意

自然言語推論を計算機で解かせるためには数十万規模のデータセットが必要となる。これまでに構築された自然言語推論データセットは言語間でその規模に大きな隔たりがある。研究が盛んな英語では、SNLI (Stanford Natural Language Inference) [1] や MultiNLI (Multi Natural Language Inference) [2] など、数十万規模の問題からなるデータセットが整備されている。一方、英語以外の言語では、大規模なデータセットは存在しない。例えば、日本語では、現存する自然言語推論のデータセットにはたかだか数千件の問題 [3] [4] しか収録されていない。これは、自然言語推論の研究の進展に大きな影響を与えている。

このような背景から、本研究では、機械翻訳に基づく、安価かつ高速な自然言語推論データセットの構築手法を提案する。一般に、自然言語推論データセットの構築は、問題の作成と検証の二つのステップからなり、この手続きによって数十万規模の問題を作成しようとすると、莫大な費用と長い開発期間を要する。本研究では、英語ですでに構築されているデータセットを自動的に翻訳し、フィルタリングを行うことによって、これらの課題を回避しつつ、高品質なデータセットの構築する。

提案する構築手法は二つのステップからなる。まず、英語のデータセットを機械翻訳によって目的の言語に変換する。次に、翻訳によって生じるノイズを軽減するため、フィルタリングを行う。フィルタリングの手法として、評価データと学習データに対し、それぞれ別のアプローチをとる。評価データは、精度を正確に測るため、クラウドソーシングを用い、人手でフィルタリングを行う。フィルタリングは、翻訳後の文が目的の言語において不自然な文を取り除き、さらに、ラベルが変化する例を取り除く。学習データは、大規模な自然言語推論データセットでは数十万の規模の問題が存在し、クラウドソーシングでフィルタリングを行うと莫大な費用がかかってしまう。したがって、計算機により自動で、効率良くデータをフィルタリングする。本研究では、機械翻訳を用いた逆翻訳による手法と、言語モデルによる手法の二つを提案する。

逆翻訳による手法では、まず、目的の言語に翻訳された文を機械翻訳によって元の言語に翻訳し直す。そして、元の文との類似性をスコア化し、そのスコアを基にデータをフィルタリングする。言語モデルによる手法では、目的の言語における言語モデルを作成し、翻訳後の文が目的の言語において自然な文となっているかを数値化し、フィルタリングを行う。

これらの手法により構築したデータセットを用いて自然言語推論モデルを学習し、その精度を測ることによって、データセットの品質の検証を行う。自然言語推論モデルとして、本研究では目的の言語で pre-training された汎用言語モデルを、自然言語推論タスクに fine-tuning することで学習する。

本研究では SNLI を翻訳の対象とし、目的の言語は日本語として実験を行った。SNLI は、現存する最大級の自然言語推論データセットであり、標準的なデータセットの一つとして用いられている。本研究の提案手法を用いることにより、評価データを 3,917 ペア、学習データを 53 万ペア構築した。このデータセットは、自然言語推論モデルにより 93.0%の精度で解くことが可能である。

2. 関連研究

本節では大規模な自然言語推論データセットと自然言語推論が多言語に応用された研究について述べる。

2.1 自然言語推論データセットの構築

現在、英語では、SNLI [1] や MultiNLI [2] など、数十万規模の問題からなるデータセットが整備されている。

SNLI は、Stanford 大学が開発した現存する最大級の自然言語推論データセットで、標準的なデータセットとして用いられている。このデータセットには評価データと開発データが 1 万ペアずつ、学習データが 55 万ペア存在する。SNLI に収録されている問題の一例を示す。

(2) a. 前提：An older man is drinking orange juice at a restaurant.
b. 仮説：A man is drinking juice.
c. ラベル：entailment

SNLI は、全て写真の説明文を元に作成しているため、視

[1] 京都大学
Kyoto University
[†1] 現在，早稲田大学
Presently with Waseda University
[a)] takumiyoshiko@nlp.ist.i.kyoto-u.ac.jp
[b)] dkw@waseda.jp
[c)] kuro@i.kyoto-u.ac.jp

覚的な情報を元にした文が多い。そもそも、自然言語推論は自然言語理解を目的としたタスクであり、その観点からすると、SNLI は自然言語推論のデータセットとしてはカバレッジが不十分なものである。

そのような欠点を解消するために作られたのが MultiNLI である。MultiNLI は様々なジャンルで用いられる言葉を用いて、会話、手紙、電話、雑誌、フィクション、ノンフィクションなど全部で 10 種類のジャンルについてデータセットを作成している。そのデータの有用性から、言語モデルの性能を評価するためのベンチマークの一つである GLUE [5] に採用されている。MultiNLI のデータセットサイズは評価データと開発データが 2 万ペアずつ、学習データが 39 万ペアである。

2.2 自然言語推論の多言語への応用

自然言語推論データセットは、言語間でその規模に大きな隔たりがある。SNLI、MultiNLI はどちらも英語のデータセットであるが、それ以外の言語では、自然言語推論のデータセットは不足しているというのが現状である。数十万規模の自然言語推論データセットを一から構築することは非常に困難であり、現実的ではない。

こうした背景から、英語の大規模なデータセットを用いて、自然言語推論を他の言語で解く手法がいくつか考案されている [6]。一つ目は、英語の学習データ全てを目的の言語に翻訳し、その目的の言語で自然言語推論モデルを学習させる手法である。二つ目は、評価データを英語に翻訳し、元の英語の学習データで学習させたモデルを用いて自然言語推論を解く手法である。三つ目は、Multi-lingual Sentence Encoder を用いた手法である。Multi-lingual Sentence Encoder は、多言語に対応した言語埋め込みを行うことで、多言語に対応したエンコーダーを学習させる。これにより、翻訳に頼ることなく英語の学習データでモデルを学習し、目的の言語で自然言語推論を解くことが可能となる。

これらの手法によるモデルの精度を測るため、用いられているデータセットが XNLI [6] である。XNLI は、MultiNLI と同様にクラウドソーシングを用いた手法により英語の評価データセットを構築し、プロの翻訳者によって、フランス語、ドイツ語、スペイン語、中国語、スワヒリ語など、リソースの少ない言語も含め 15 言語に翻訳したデータセットである。モデルの学習データとしては、MultiNLI と同じデータを用いる。

3. 自然言語推論データセットの多言語化手法

自然言語推論データセットの多言語化に当たり、まず、既存の大規模な自然言語推論データセットを目的の言語に機械翻訳する。その後、評価データと学習データでそれぞれ別々の手法でフィルタリングを行う。評価データは正確性が求められるため、クラウドソーシングを用いて人手で正確に行う。また、ラベルの正しさについても検証を行い、より正確を期す。学習データはデータサイズが大きく、クラウドソーシングを用いると莫大な金額がかかるため、自動で効率的にフィルタリングする。

3.1 機械翻訳

まず、英語のデータセットを目的の言語に翻訳する。数十万ペアのデータセットを人手で翻訳するのは莫大なコストがかかるため、本研究では機械翻訳を用いる。機械翻訳の手法として本研究では Google 翻訳 [7] を用いる。安価かつ、多言語に対応しており、高精度な翻訳を行えるためである。

3.2 評価データのフィルタリング

機械翻訳を行うと、文として意味の通らないものや、元の文とは違った意味の文が生成される可能性がある。評価データに、これらのノイズが含まれていると、モデルの評価を正しく行うことができなくなる。

評価データを正確に構築するため、クラウドソーシングを用いてノイズをフィルタリングする。フィルタリングは翻訳誤りのフィルタリングとラベル変化のフィルタリングの二段階に分けて行う。翻訳誤りのフィルタリングは、文として意味の通らないもののフィルタリングを行う。ラベル変化のフィルタリングでは、翻訳の際に元の文と意味が変化したことによってラベルが変化したデータをフィルタリングする。

3.2.1 翻訳誤りのフィルタリング

まず、翻訳文が翻訳された言語として意味の通らない文をフィルタリングする。本研究では、クラウドソーシングを用いる。

手法としては、5 人のクラウドワーカーに翻訳文を提示し、翻訳された言語として意味が通るか、通らないかの二択で回答させる。5 人中 3 人以上が意味が通ると答えたデータを翻訳が成功したものとし、それ以外のデータを翻訳が失敗したものとする。元の文はクラウドワーカーには見せず、翻訳された文のみを提示する。そのため、文の意味が変化しているかどうかはこの段階では考慮しない。

3.2.2 ラベル変化のフィルタリング

翻訳誤りのフィルタリングを行った後、前提と仮説の二つの文がどちらも翻訳に成功したペアだけを抽出する。翻訳誤りのフィルタリングでは、翻訳による意味の変化に対しては考慮していない。したがって、このままデータを作成するとラベルが元のラベルに対して変化してしまう場合も考えられるため、ラベル変化のフィルタリングを行う。

手法としては、まず、5 人のクラウドワーカーに前提と仮説の二つの文を提示し、これらの文の関係が含意、矛盾、中立のうちどれかを三択で答えてもらう。3 票以上集まったラベルを採用し、どのラベルにも 3 票以上集まらなかっ

たデータは取り除く。そして、ラベルの変化した例を取り除くため、採用されたラベルと元のラベルが一致したものだけを評価データとして採用する。

3.3 学習データのフィルタリング

大規模な自然言語推論データセットにおいては、学習データは数十万規模の問題が存在する。したがって、安価にフィルタリングを行うためには、計算機で自動的に翻訳文を評価する必要がある。

本研究では、翻訳文のフィルタリングの手法として、逆翻訳による手法、言語モデルによる手法の二つを提案する。

3.3.1 逆翻訳によるフィルタリング

3.1 節で翻訳したデータのうち、学習データのみについて逆翻訳を行い、BLEU によって元の文との類似度を比較する。逆翻訳は 3.1 節と同様に Google 翻訳を用いて行う。

BLEU [8]（BiLingual Evaluation Understudy）は、機械翻訳されたテキストを自動的に評価するための指標である。BLEU によって元の文と逆翻訳された文を比較し、その一致率を測ることによって、間接的に翻訳された文の評価を行う。ここでは元の英文を参照訳、逆翻訳された文を翻訳文とする。BLEU は以下のように定義される。

$$BLEU = BP_{BLEU} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

ただし、

$$\mathrm{BP}_{BLEU} = \begin{cases} 1 & (c \geq r) \\ \exp(1 - \frac{r}{c}) & (c < r) \end{cases}$$

$$w_n = 1/N$$

$$p_n = \frac{\text{翻訳文と参照訳で一致した n-gram 数}}{\text{翻訳文の中の全 n-gram 数}}$$

r は参照訳の長さ、c は翻訳文の長さである。p_n は、評価コーパス全体について、翻訳文と参照訳を比較し、n-gram の一致度を算出しているものである。これを 1-gram から N-gram について幾何平均を求めることにより、スコアを算出する。本来 BLEU はコーパスベースの指標であり、通常は $N = 4$ が用いられるが、個々の短い文を評価しようとすると 4-gram における p_n がほとんど 0 となる。よって本研究では $N = 1$ として計算を行う。

この計算式を用いて元の文と逆翻訳された文の BLEU スコアを出力し、適当な閾値を設定することでデータのフィルタリングを行う。

3.3.2 言語モデルによるフィルタリング

翻訳文の文としての自然さを数値化してフィルタリングするため、本研究では、LSTM [9] を用いて目的の言語における言語モデルを学習する。入力は文を単語ごとに区切ったものを埋め込んだものである。出力はクロスエントロピーを単語数で割った値とする。文として不自然であると予測するほど高い値を出力し、文として自然であるほど低い値を出力する。この出力を元にデータのフィルタリングを行う。

4. 日本語データセットの構築

データセット作成にあたって、本研究は元データとして SNLI を使用し、日本語を目的の言語とする。

SNLI は視覚的な情報に基づいた文が多く、カバレッジが不十分であるが、現存する自然言語推論データセットの中で最大級のデータセットサイズを誇り、依然として標準的な自然言語推論データセットの一つであるため採用した。

対象言語を日本語としたのは、大規模な自然言語推論データセットが存在しないことと、自然言語推論データセットを日本語に翻訳した前例がないためである。

4.1 SNLI の英日機械翻訳

まず、SNLI のデータセット 57 万ペアから前提文と仮説文を全て抽出しまとめて翻訳を行った。全部で 57 万ペアあるため前提と仮説合わせて 114 万文存在するが、SNLI ではデータセット構築の際にクラウドワーカーに 1 つの前提文に対して 3 つの仮説文を書かせているため実際には前提文は少なくとも 3 文は重複が存在する。重複した文を取り除くと翻訳すべき文は全部で 65 万文となる。それらすべての文を Google 翻訳を用いて、英日機械翻訳を行った。翻訳にかかった費用は 63,196 円である。下に翻訳文の成功例と失敗例を 2 つずつ示す。a が元の文、b が翻訳後の文である。

(3) a. A man is playing with friends.
b. 男は友達と遊んでいます。

(4) a. A woman is eating pasta in a restaurant.
b. 女性がレストランでパスタを食べています。

(5) a. A woman having a cavity removed.
b. 空洞が除去された女性。

(6) a. A skinny man in a tank top watches TV inside.
b. タンクトップの skin せた男が中のテレビを見ています。

例 (3) (4) を見ると平易な文はある程度正確に訳せていることが分かる。例 (5) を見ると「cavity」という単語が本来「虫歯」と訳されるべきところが誤って「空洞」と訳されている。このように複数の意味を持つ英単語は誤って翻訳される可能性がある。また、例 (6) をみると「skinny」は本来「痩せた」の意味で使われているが、翻訳の際に「skin」＋「ny」という形で分解されてしまい、「skin」という単語が翻訳されずにそのままの形で出力されてしまった。このように英単語の形のまま出力されてしまう例は所々見受け

られた。

翻訳された文をランダムに 100 文抽出して日本語として意味が通るかを人手で検証した。翻訳文が意味の通る文になっていた割合は 78 %であった。

4.2 評価データのフィルタリング

3.2 節で提案したように、評価データのフィルタリングは、翻訳誤りのフィルタリングとラベル変化のフィルタリングの 2 段階に分けて行った。

4.2.1 翻訳誤りのフィルタリング

評価データの翻訳文の検証は Yahoo!クラウドソーシングを用いて行った。検証を行ったデータ数は 10,000 ペア (13,159 文) であり、かかった金額は 65,800 円である。 3.2.1 の手法により行った。以下にクラウドワーカーへの質問の例を示す。

質問： 提示された文が日本語として意味が通るかどうかについて以下の 2 つの選択肢から選んでください。
提示文： 数人の子供が森で遊んでいます
選択肢： (1) 日本語として意味が通る
(2) 日本語として意味が通らない
想定される回答： (1) 日本語として意味が通る

検証の結果としては翻訳文 13,159 文のうち翻訳が成功したデータは 10,295 文 (78.2 %)、翻訳が失敗した例は 2,864 文 (21.8 %) であった。この翻訳が成功した割合は機械翻訳後に人手で翻訳文の検証を行ったときの割合 (78 %) とかなり近い値を示しており、クラウドソーシングの結果がある程度信頼できるものと考えられる。

この結果を元に、前提文と仮定文がともに翻訳に成功したペアのみを採用することにし、その結果 10,000 ペア中残った評価データは 5,474 ペアとなった。

4.2.2 ラベル変化のフィルタリング

まず、翻訳によってラベルが変化してしまう場合について分析した。英語から日本語への翻訳の例を挙げる。

前提：An old man with a package poses in front of an advertisement.
仮説：A man poses in front of an advertisement.
ラベル：含意
↓
前提：パッケージを持つ老人が広告の前でポーズします。
仮説：男が広告の前でポーズをとる。
ラベル：中立

図 1 ラベル変化の例

図 1 では、翻訳前の前提文の「An old man」の部分が翻訳後では「老人」と訳されている。翻訳としては間違っていないが、翻訳前には性別が男として特定されていたものが、「老人」という性別の特定されていない表現に変わったため、ラベルが含意から中立に変わってしまう。

この例を見ると分かるように日本語として意味が通るように翻訳されていたとしても微妙な表現の違いやニュアンスの違いでラベルが変化してしまう場合がある。そこで、正確性が必要な評価データは、ラベルが変化する可能性を考慮するためにクラウドソーシングを用いてラベルをフィルタリングする。

翻訳誤りのフィルタリングを行った後、前提と仮説の 2 つの文がどちらも翻訳成功したペア (5,474 文) だけを抽出してラベルの検証を行った。ラベルの検証は翻訳文の検証のときと同様に Yahoo!クラウドソーシングで行った。かかった金額は 27,400 円である。5 人のクラウドワーカーに (A) 前提文と (B) 仮説文の 2 つの文を見せ、これら 2 つの文の関係が含意、矛盾、中立のうちどれかを 3 択で答えてもらった。

以下にクラウドワーカーへの質問の例を示す。

質問： (A) と (B) は同時に起こる出来事と仮定します。A と B の関係について答えてください。
提示文： (A) 2 匹の犬が雪の中で立っています。
(B) 2 匹の犬がプールで泳いでいます。
選択肢： (1) A が正しいなら B も正しい
(2) A が正しいなら B は正しくない
(3) A が正しいとしても B の真偽は分からない
想定される回答： (2) A が正しいなら B は正しくない

3 票以上集まったラベルを採用し、どのラベルにも 3 票以上集まらなかったデータは取り除いた。そして、採用されたラベルと元のラベルが一致したものだけを評価データとして採用した。この手法で実際に評価データとして有効なものは 5,474 文中 3,917 文であった。

4.3 学習データのフィルタリング

学習データに関しては 55 万ペアあり、クラウドソーシングで検証を行うと莫大な金額がかかってしまうため、機械を用いて翻訳文を評価し、ある一定の閾値でフィルタリングを行うことで効率良く検証を行う。本研究では逆翻訳による手法と LSTM 言語モデルによる手法の 2 つで実験する。

4.3.1 逆翻訳によるフィルタリング

翻訳したデータのうち、学習データのみを取り出して、Google 翻訳により日英翻訳を行った。かかった費用は 19,634 円である。

逆翻訳された文を、元の文と BLEU の計算式に基づいて算出したスコアを元にデータのフィルタリングを行う。BLEU のスコアは 0〜1.0 の値をとり、1.0 に近いほど翻訳の精度が高いと考えられる。本研究では、閾値を 0.2〜0.6 まで 0.1 刻みで変化させてフィルタリングを行った。

4.3.2 LSTM 言語モデルによるフィルタリング

言語モデルによる翻訳文評価に取り組むため、本研究では、Wikipedia の日本語の全記事を元に、LSTM 言語モデ

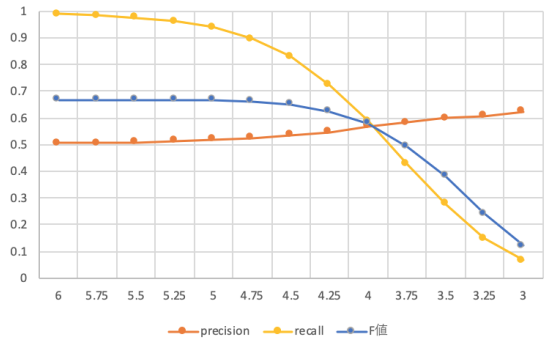

図 2　LSTM スコアの閾値と precision、recall、F 値

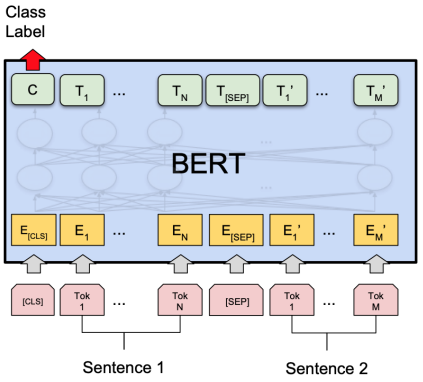

図 3　BERT を用いた自然言語推論モデル ([10] より)

ルを作成した。

まず、この LSTM 言語モデルのフィルターとしての性能を検証するために、 4.2 節で検証した翻訳文を使って実験を行った。まず、翻訳が成功したデータを正解ラベル、失敗したデータを誤りラベルとする。本実験では、結果がわかりやすいように元のデータから正解ラベルと誤りラベルの数が同じになるようそれぞれ 2,864 文ずつ抽出した。これらの翻訳文を LSTM を使ってスコアを出力し、そのスコアに対して閾値を設定し、閾値を超えたデータを取り除いてデータをフィルタリングすると precision、recall、F 値がどのように変化するかを検証した。

結果として図 2 のように、閾値を低く設定するほど、precision は上がる。一方で、recall は下がるため、F 値としては閾値 4.5 付近まではほぼ横這いであった。

学習データの前提、仮説の全ての文のスコアを、言語モデルを使用し算出する。そして、ある閾値を設定しその閾値を超えるスコアのデータを取り除いていく。閾値は 5.0 から 3.5 までの間で 0.25 刻みで変化させてフィルタリングを行う。

5. 自然言語推論の日本語モデルの学習

構築したデータセットの有用性を検証するため、BERT [10] による自然言語推論モデルを学習し、評価する。また、翻訳された学習データをフィルタリングし、BERT の評価値がどう変化するかを検証する。

5.1 BERT に基づく自然言語推論モデル

BERT (Bidirectional Encoder Representations from Transformers) は self-attention 機構を利用した Transformer [11] をベースとしたモデルであり、pre-training と各タスク固有の fine-tuning を施すことで，様々なタスクで state-of-the-art を達成している。Pre-training では Masked Language Model と Next Sentence Prediction という 2 つのタスクで学習を行う。

本研究では日本語で Pre-training 済みのモデル [12] を自然言語推論タスクに fine-tuning する。BERT の Transformer の上にタスク固有の最終層を加え、Transformer の隠れ層のベクトルをこの最終層に入力する。この最終層の出力がタスク固有の正解を予測できるように、タスクごとのデータセットで BERT モデルのパラメータと、最終層の重みを学習する。本研究では自然言語推論を解かせるため、図 3 のように入力を [CLS] (文 1) [SEP] (文 2) の構成にし、[CLS] トークンに 3 値分類の値、文 1 に前提、文 2 に仮説を入力する。最終層の出力を、含意、矛盾、中立の 3 値分類の値をソフトマックスにより算出する。

5.2 実験・結果

データをフィルタリングした後、前提、仮説がともに残っていた場合のみ、そのペアを使用することにする。そうして作成した学習データを 5.1 節の自然言語推論モデルに学習させ、4.1 節で作成した評価データを使って評価値を計算する。閾値を変化させることでモデルの評価値が上がるどうかを検証する。

まず、学習データを全くフィルタリンをせずに学習を行ったとき、評価値は 0.929 となった。

逆翻訳によるフィルタリングでは、閾値は 0.2 から 0.1 ずつ値を変えていき、データのフィルタリングを行った。その結果、閾値を 0.2 にしたとき自然言語推論モデルの評価値は減少したが、0.3 にしたときは 0.2 のときに比べ少し上昇した。さらに閾値を上げていくと、それ以降はモデルの評価値は減少していった。全体としては、フィルタリングをする前よりも大きく評価値が上がることはなかった。

LSTM のスコアの閾値を 5 から 3.5 まで、0.25 ずつ下げていき、データをフィルタリングしていった結果、閾値を下げるほど、自然言語推論モデルの評価値は減少していくという結果になった。フィルタリングをする前よりも評価値が上がることはなかった。

5.3 議論

学習データのフィルタリングを行う前の評価値は 0.929 となった。この数値は、元の SNLI の精度の最高値である

閾値	-	0.1	0.2	0.3	0.4	0.5	0.6	0.7
データサイズ	55 万	53 万	51 万	47 万	41 万	31 万	20 万	10 万
BERT の評価値	0.929	0.930	0.924	0.926	0.920	0.915	0.904	0.887

図 4 BLEU スコアの閾値とモデルの評価値

閾値	-	5.25	5.0	4.75	4.5	4.25	4.0	3.75	3.5
データサイズ	55 万	49 万	46 万	42 万	36 万	28 万	20 万	11 万	5.2 万
BERT の評価値	0.929	0.925	0.925	0.923	0.922	0.913	0.903	0.892	0.871

図 5 LSTM スコアの閾値とモデルの評価値

91.6%を上回っている。これは図 1 のようなラベルの変化を伴うような難しい問題を削除したためだと考えられる。

まず、実際にこのモデルが正解した例と誤答した例について考察する。

表 1 に正解した例を示す。表 1(1) では、仮説の「サッカーの試合で得点しようとする」の部分が、前提の「サッカーの試合中にゴールに向かってサッカーボールを蹴ります」の言い換えになっており、正解は含意である。モデルも含意と判定しており、文の言い換えを正しく推論できていることがわかる。また、表 1(2),(3) のような矛盾、中立の問題も正しく解くことができている。

表 2 に誤答した例を示す。表 2(1) では、仮説の「暖かい」の部分は前提には明示されていないが、冬服は冬に着るため暖かいものであり、正解は含意である。こうした推論がモデルでは行われず、文字通りの意味だけで、中立という誤ったラベルを判定してしまったと考えられる。表 2(2) では、前提では女性の服装について言及しているのに対し、仮説では女性たちの行動について説明しており、それぞれ別々の事柄について言及しているが、矛盾はしていないため正解は中立である。一方で、モデルは矛盾と判定しており、このような矛盾はしていないがそれぞれ別々のことについて言及しているような問題を矛盾と間違える例は他の問題でもいくつか見られた。

逆翻訳によるフィルタリングでは、BLEU のスコアが 0.1 のとき、最高値 0.930 となり、フィルタリングを行う前と比べて大きく精度が上がることはなかった。これは翻訳の精度が上がる影響よりもデータ数が少なくなる影響をより強く受けていると考えられる。本研究においては、前提文と仮説文のスコアの両方が閾値以下であるデータのみを採用しており、SNLI のデータセットは 1 つの前提文に対して仮説文が複数存在するという構造になっているため、1 つの前提文を取り除くとそれに付随して複数のペアが取り除かれしまう。したがって、一つの前提文を取り除くために少なくとも 3 ペア、多いものでは 15 ペアものデータが取り除かれてしまう。このように一文を取り除くために多くのデータが取り除かれてしまうためモデルの評価値が上がらなかったと推測される。

LSTM 言語モデルによるフィルタリングの結果、閾値を低くすればするほどフィルタリング後の自然言語推論モデルの評価値は減少していった。これは逆翻訳によるフィルタリングと同様に、翻訳の精度が上がる影響よりもデータ数が少なくなる影響をより強く受けていると考えられる。

6. おわりに

本研究では機械翻訳による自然言語推論データセットの多言語化を目的と置いたが、機械翻訳による言語の変換により、日本語の自然言語推論データセットを構築することに成功した。このデータセットは BERT による自然言語推論モデルを用いて 93.0 %の精度で解くことが可能である。数十万規模の自然言語推論データセットを、英語から日本語に変換させた例は存在せず、日本語では初の試みである。

本研究の提案手法は、言語モデルを構築できる程度の言語資源を持つ言語であれば、どの言語にでも適用できる。したがって、言語資源が比較的少ない言語であっても、数十万規模の自然言語推論データセットの構築が可能となる。

今後は中国語など、他言語にもこの手法を適用して実験を行っていく。また、省略解析などの言語理解タスクと併用してマルチタスク学習に適用することで、自然言語推論モデルのさらなる精度向上を目指していく。

	正解ラベル	予測ラベル	前提	仮説
(1)	含意	含意	１３番は、子供のサッカーの試合中にゴールに向かってサッカーボールを蹴ります	サッカーの試合で得点しようとするプレーヤー
(2)	矛盾	矛盾	男と女が公園で話している	男と女が水族館で話しています。
(3)	中立	中立	黄色の救助用具に身を包んだ男が野原を歩いています。	男は誰かを探しています。

表 1 モデルが正解した問題の例

	正解ラベル	予測ラベル	前提	仮説
(1)	含意	中立	冬服を着た３人の子供が荷物を押しながら森の中を歩いています。	３人の子供は暖かい冬服を着ています。
(2)	中立	矛盾	２人の女性、１人はスカーフ、もう１人は広いつばの帽子をかぶっています	２人がテーブルでパン生地を叩いています。
(3)	中立	含意	番号９１６は、彼がレースに勝つことを望んでいます。	人は彼がレースに勝つと賭けている。

表 2 モデルが誤答した問題の例

参考文献

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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. 1112–1122. Association for Computational Linguistics, 2018.

[3] 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫. 日本語 textual entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会 第 14 回年次大会, pp. 1140–1143, 2008.

[4] 川添愛, 田中リベカ, 峯島宏次, 戸次大介. 日本語意味論テストセットの構築. 言語処理学会 第 21 回年次大会, pp. 817–820, 2015.

[5] 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, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[6] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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. Association for Computational Linguistics, 2018.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, 2016.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, USA, 2002. Association for Computational Linguistics.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, pp. 1735–1780, 12 1997.

[10] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, 2017.

[12] 柴田知秀, 河原大輔, 黒橋禎夫. Bert による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会, pp. pp.205–208, 2019.