日本語 WiC における半教師あり学習の単語頻度に応じた性能評価

山下莉実[1]　佐々木稔[2]
[1]茨城大学工学部情報工学科 [2]茨城大学理工学研究科 情報科学領域
{22t4078n}@vc.ibaraki.ac.jp
{minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

本研究では,日本語 WiC タスクにおいて,半教師あり学習モデル COSINE の性能を,単語の出現頻度に基づいて分析した.対象単語を頻度別に分割し,疑似ラベル採用時の確信度閾値を変化させた場合の挙動を,教師あり学習（SL）と比較した.その結果,中頻度語においては,COSINE が低い閾値設定で SL に見られる過学習を抑制し,文脈をより適切に捉える傾向が確認された.一方,低頻度語では,ノイズの影響で COSINE の性能が低下し,SL の方が高い正答率を示すケースも観測された.結論として,COSINE は過学習抑制を目的とした正則化的手法として有効であり,単語頻度や確信度に応じて手法や閾値を切り替えるハイブリッド戦略の有効性が示唆された.

1 はじめに

自然言語処理において,語義曖昧性解消の一形態である Word-in-Context（WiC）タスクは,文脈理解能力を評価する重要なベンチマークとして注目されている[1].WiC は,同一単語を含む 2 文が与えられ,その単語が同一の意味で使用されているかを判定する二値分類タスクである.

WiC データセットの大きな課題の一つに,学習データの確保が困難であるというロングテール問題がある.言語内には出現頻度の低い単語が膨大に存在しており,これらすべてに対して十分な教師ラベルを付与することは困難である.このような少量のラベル付きデータと大量のラベルなしデータを組み合わせて活用する手法として,半教師あり学習の有効性が期待されている[2].半教師あり学習において,学習の質を左右するのは,モデルが生成する疑似ラベルの信頼性である.しかし,単語の出現頻度の違いや,疑似ラベルを採用する際の確信度の閾値設定が,最終的なモデルの挙動や文脈理解力にどのような影響を与えるかについては,詳細な分析がなされていない.

そこで本研究では,日本語 WiC タスクにおいて,対照学習と自己学習を組み合わせた半教師ありモデルである COSINE[4]と教師あり学習[3]の比較評価を行う.具体的には,対象単語をコーパス内の出現頻度に基づき 4 つのグループに分割し,それぞれの頻度帯における性能を検証する.さらに,疑似ラベル採用時の閾値を変動させることで,モデルの過学習の抑制と学習バイアスの相関を詳細に分析し,低頻度語を含む多様なデータに対する最適な学習戦略を明らかにすることを目的とする.

2 関連研究

2.1 教師あり学習を用いた WiC タスク

Pilehvar ら[3] は,BERT などの事前学習済み言語モデルを文脈判定タスク用にファインチューニングする標準的な教師あり学習手法を確立した.本研究における教師あり学習（SL）もこの手法に基づいている.

2.2 半教師あり学習を用いた WiC タスク

学習データが極端に少ない状況下において,半教師あり学習のアプローチが研究されている.例えば Schick ら[2] は,パターン活用学習 PET とラベルなしデータの活用を組み合わせる iPET 手法を提案し,わずかな教師ありデータしか利用できない WiC タスクにおいても,モデルの性能を大幅に改善できることを実証した.

2.3 COSINE

半教師あり学習モデル COSINE[4]は,事前学習済み言語モデルを,正解ラベルのないデータ,またはノイズを含んだ疑似ラベルのみで効果的にファインチューニングすることを目的とする.従来の自己学習や弱教師あり学習では,ラベルに含まれるノイズに

対してモデルが過学習するという問題があった.これに対し,COSINE は,対照学習（Contrastive Learning）と自己学習（Self-training）を組み合わせることで,ノイズの影響を抑制しながらモデルの性能を向上させる.

3 分析手法

3.1 WiC データセット

Raganato ら[5]により提案された,WiC の多言語データセット XL-WiC の日本語サブセットを使用した.データは github の公式リポジトリ[6]より取得した.各データは以下の前処理を行い,JSONL 形式に整形した.

{target_word：対象単語,
sentence1/sentence2: 対象単語を含む 2 つの文
label: 正解ラベル（True or False）,
start1/start2：各文における対象単語の開始位置インデックス
end1/end2：各文における対象単語の終了位置インデックス
rule_label：疑似ラベル
（ニューラルネットワークにより付与された予測ラベル）（True or False）}

元データにはない疑似ラベルは,XLM-RoBERTa-large により抽出した対象語埋め込みおよび文埋め込みを入力とする二値分類ニューラルネットワークを用いて生成した.

成形したデータを以下の 4 つのサブセットに分割した.

- 訓練データ (Train): 教師あり学習,および COSINE の教師モデルの学習に使用
- 評価データ (Dev): モデルのハイパーパラメータ調整および学習中の評価に使用.
- テストデータ (Test): 最終的な性能評価に使用.
- ラベルなしデータ (Unlabeled): 半教師あり学習 COSINE において使用.元データの正解ラベル”label”は無視し,教師モデルが付与した疑似ラベル”rule_label”のみを学習に利用する.

3.2 実験モデル

ベースモデルとして RoBERTa-base を採用し,以下の 2 つの手法で WiC タスクを実行した.

3.2.1 教師あり学習(SL)手法

ラベル付きの訓練データのみを用いてモデルを学習させる.ベースライン性能および学習傾向を確認する.

3.2.2 半教師あり学習 (COSINE)手法

自己訓練（Self-Training）と対照正則化（Contrastive Regularization）を組み合わせた手法である COSINE を用いる. フレームワークを図 1 に示す.

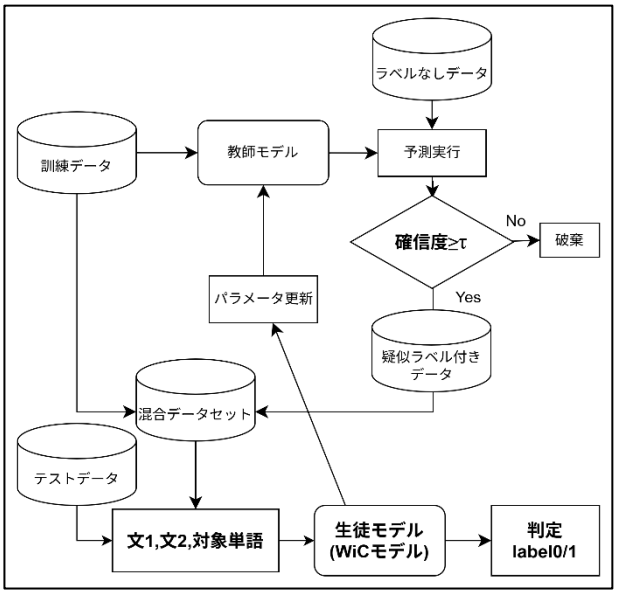

図 1 COSINE のフレームワーク

図 1 に示したモデルの動作について,以下で説明する。

step1 訓練データのみで教師モデルを初期学習させる.

step2 教師モデルを用いて,ラベルなしデータに対し予測を行い,確信度とともに疑似ラベルを付与する.閾値 τ を超える信頼性の高い疑似ラベル付きデータを採用する.

step3 ラベル付きの訓練データと疑似ラベルデータを混ぜて生徒モデルを学習させる. モデルは,「対象単語, 文 1,文 2」を受け取り, 同じ意味であれば label0,違う意味であれば label1 を出力する.

学習した生徒モデルが次の教師モデルとなり, プロセスを繰り返す.

疑似ラベルを採用する際の確信度の閾値を 0.3 から 0.9 の範囲で変動させ,ノイズの影響とモデルの挙動の変化を検証した.

3.3 性能評価

モデルの性能は,以下の観点から評価した.

3.3.1 定量指標正答率 (Accuracy / Micro F1)

モデルが対象単語の「同じ意味（Label 1）」と「異なる意味（Label 0）」を総合的にどの程度正しく識別できているかを測定するため, 正答率（Accuracy）および MicroF1 スコアを用いる. これにより, 手法間および各閾値設定における全体的な有効性を比較する.

3.3.2 単語頻度別分析

モデルの得意不得意を詳細に分析するため,テストデータの対象単語をコーパス内の出現頻度に基づき,以下の 4 つのグループに分割して評価を行った. 頻度は現代日本語書き言葉均衡コーパス BCCWJ[7] より取得した.

- 高頻度語 (Rank 1-50): モデルが事前学習で頻繁に接している単語群.
- 中頻度語 (Rank 50-100): 一般的な単語群.
- 中低頻度語 (Rank 100-150): 中頻度語よりやや使用頻度が低い単語群.
- 低頻度語 (Rank 150-200): 学習データに含まれることが稀なロングテール領域の単語群.

4 結果

実験結果は,表 1 の通りである. この表は,単語の出現頻度ごとに,COSINE の疑似ラベル採用時の確信度閾値（0.3〜0.9）と教師あり学習の正答率を比較したものである.

4.1 性能の比較

教師あり学習（SL）と半教師あり学習（COSINE）の全体的な性能を比較した結果, COSINE は特に閾値 0.4 付近において教師あり学習を上回る正答率を示した.教師あり学習では,異なる意味（Label 0）の事例を同じ意味（Label 1）と誤分類する傾向が観測されたのに対し,COSINE は疑似ラベルを通じて否定例を学習することで,この傾向を抑制したと考えられる.

表 1　教師あり手法(SL), 半教師あり手法(COSINE)の正答率

	閾値							SL
頻度	0.3	0.4	0.5	0.6	0.7	0.8	0.9	-
1-50	0.54	0.58	0.62	0.6	0.54	0.6	0.48	0.54
50-100	0.66	0.74	0.6	0.6	0.6	0.46	0.42	0.5
100-150	0.66	0.72	0.64	0.7	0.66	0.66	0.64	0.6
150-200	0.72	0.54	0.64	0.62	0.68	0.76	0.46	0.74

4.2 単語頻度ごとの挙動

テストデータを単語の出現頻度順に 4 つのグループに分割して分析した結果,頻度帯によって有効なモデルが異なることが明らかになった.

高頻度語（1–50 位）モデルは事前学習の知識により同じ意味であるという強い確信を持っており,SL は文脈が明らかに異なる場合でも「同じ」と判定する傾向があった.COSINE は,ラベルなしデータから異なる意味を学習し,一部の単語で誤検知を抑制したが,全体的なスコアの上昇に貢献しなかった.

中頻度語（50–150 位）この領域では最も COSINE の効果を発揮した.SL がデータ不足による過学習を起こす一方で,COSINE は低閾値でノイズを許容しつつ外部データを取り込むことで,文脈の違いを正しく検知できた.しかし,閾値設定によってはノイズの影響を強く受け,ラベルが偏るという不安定さも見られた.

低頻度語（150–200 位）この領域では,COSINE の性能が低下した.教師なしデータに含まれるノイズの影響で確信度が低下し,ほとんどの単語に違うと判定する消極的な挙動が見られた.一方で,SL は同じ意味であるだろうというバイアスによって Label 1 の正解を多く拾い,高い F1 スコアを維持した.

4.3 閾値による影響

COSINE における疑似ラベル採用の閾値を変動させたところ,特に低頻度語において劇的な挙動の変化が観測された.

高閾値（0.8,0.9）ノイズを遮断するため,SL の Label 1 へのバイアスが維持されるか,データ枯渇により全て 0 と判定する状態となった.

中閾値（0.5〜0.7） 中頻度,中低頻度語においては,ノイズが中途半端に混入し,モデルに迷いが生じた結果,正答率が低下する谷間が形成された.

低閾値（0.3,0.4）弱い信号も全て取り込むことで,再び SL の Label 1 へのバイアスが強化され,見かけ上のスコアは回復した.

5 考察

5.1 文脈理解とバイアス

本実験の結果,モデルが真に文脈を読み取って判定している事例は,主に高頻度から中頻度の領域に分散して確認された.

一方で,低頻度語において観測された高い F1 スコアは,文脈理解によるものではなくデータの分布がモデルのバイアスと合致した結果である可能性が高い.低頻度語は高頻度語に比べて語義の数が少なく,特定の用法で固定される傾向がある.そのため,学習データに Label 1 が多いという偏りが生じている.また,学習データが少ない状況下では,モデルは複雑な文脈の違い(Label 0)を十分に学習しきれず,「単語が一致しているなら Label 1 である」というルールに過剰適合した可能性がある.これらが合致したことにより,数値上の性能が押し上げられたと考えられる.

5.2 ロングテール問題に対する半教師あり学習の限界

本研究の主目的であったロングテール（低頻度語）問題の解決に対し,COSINE は期待された効果を発揮しなかった.その原因として, 初期教師モデルが生成する疑似ラベルの信頼性不足が考えられる.低頻度語では文脈の多様性が十分に学習されておらず,その結果として誤った疑似ラベルが蓄積された可能性がある.

5.3 COSINE の有効性

中頻度語において SL に見られた過学習を抑制し,より慎重で汎用的な判定を可能にした点は高く評価できる.このことから,COSINE は少データ環境下で未知の知識を獲得する外部知識としてではなく,モデルの暴走を止める正則化項として機能したと結論付けられる.

6 結論

本研究では,WiC タスクにおいて,半教師あり学習（COSINE）と教師あり学習（SL）の比較を行った.実験を通して得られた主な知見は以下の通りである.

はじめに,モデル特性の違いが明らかになった.SL は未知のデータに対しても積極的に同一性を認める正解を見つける力（Recall）に優れる一方,COSINE は周辺文脈から用法の差異を厳密に識別する間違いを見抜く力（Precision）において高い能力を示した.次に,単語の出現頻度と最適な学習手法の関係性が示唆された.中〜高頻度語の領域では低閾値の COSINE が有効であり,データが欠乏している低頻度領域では SL（または高閾値の COSINE）が有効であった.さらに,疑似ラベルの閾値が持つ役割が判明した.閾値は単なるフィルタリングではなく,モデルが積極的か消極的かを決定する重要なパラメータである.

単一のモデル,単一の閾値ですべての単語に対応することは困難であると結論付けられた.今後の課題としては,単語の頻度やモデルの確信度に応じて,高頻度語には低い閾値,低頻度語には高い閾値を適用する動的閾値の導入や,データ不足を補うための辞書定義の活用が考えられる.

参考文献

[1] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato and Roberto Navigli. **Recent Trends in Word Sense Disambiguation: A Survey**. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pp4330-4338, 2020.

[2] Timo Schick and Hinrich Schütze1. **It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners**.NAACL, pp2339–2352, 2021.

[3] Mohammad Taher Pilehvar, and Jose Camacho-Collados. **WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations**. NAACL, pp1267–1273, 2019.

[4] Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, Chao Zhang. **Fine-Tuning Pre-trained Language Model with Weak Supervision: A**

Contrastive-Regularized Self-Training Approach. NAACL, pp1063-1077, 2021.

[5] Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, Mohammad and Taher Pilehvar. **XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization**. EMNLP, 2020.

[6] **XL-WiC: The Multilingual Word-in-Context Dataset**. https://pilehvar.github.io/xlwic/

[7] 国立国語研究所(2025)「現代日本語書き言葉均衡コーパス」（バージョン 2021.03,中納言バージョン 2.7.3,分類語彙表情報 2025.03）https://clrd.ninjal.ac.jp/bccwj/（2025-10-4 確認）