Weaker Than You Think: A Critical Look at Weakly Supervised Learning

Dawei Zhu Xiaoyu Shen Marius Mosbach Andreas Stephan Dietrich Klakow

2023/9/28 武田笹野研究室 M1 陽田祥平

概要

- 品質の悪いデータで学習を行う**弱教師あり学習**が過大評価されている点を指摘
- 弱教師あり学習が検証に用いるクリーンなデータで学習を行なった場合
  弱教師あり学習の手法を上回る性能を報告
- 以上の知見を基に、弱教師あり学習が有効である場面の考察や、
  より効果的な学習手法を提案

はじめに

- 機械学習において訓練データのアノテーションがコストのボトルネックとなっている
  →コストの小さい弱い教師信号で学習を行う**弱教師あり学習**が多く提案されている

・知識や経験則によるルールベース
・クラウドソーシング
など

- 一部の手法では教師あり学習の性能に匹敵するものも存在

- このような手法では最適なハイパラやモデルの選択に
  適切なアノテーションがされているクリーンな検証用データを利用
  →**この検証用データを学習に利用すればいいのでは？**

はじめに

- 以下を検証するための実験を実施

  ・弱教師あり学習と検証用データのみで学習を行なったモデルの性能を比較

  ・クリーンな検証用データが無いとき、弱教師あり学習は機能するのか

  ・検証用データの数によって弱教師あり学習の性能はどう変化するか

  ・弱教師あり学習にてクリーンなデータで追加学習した場合性能はどう変化するか

準備 - データセット

- データセットにWRENCHを利用

- 弱教師あり学習のベンチマークとして提案されている

- 4つのテキスト分類タスク(AGNews,IMDb,Yelp,TREC)、
  2つの関係分類タスク(SemEval,ChemProt)、
  2つの固有表現認識タスク(CoNLL-03,OntoNotes)からなる

- それぞれのタスクに対してラベル生成則が定義され、
  それによりラベルを決定した弱い訓練データが用意されている
  （例）IMDb…文中に特定の表現がある場合極性を決定

Label	Labeling Function
POS	beautiful, handsome, talented
NEG	than this, than the film, than the movie
POS	.*(highly|do|would|definitely|certainly|strongly|i|we).*(recommend|nominate).*
POS	.*(high|timeless|priceless|HAS|great|real|instructive).*(value|quality|meaning|significance).*

準備 - 弱教師あり学習手法

- 代表的な弱教師あり学習手法である
  L2R、**MLC**、**BOND**、**COSINE**、**MetaWN**、**Denoise**を利用
    - 主にメタ学習などで教師データのノイズを除去する

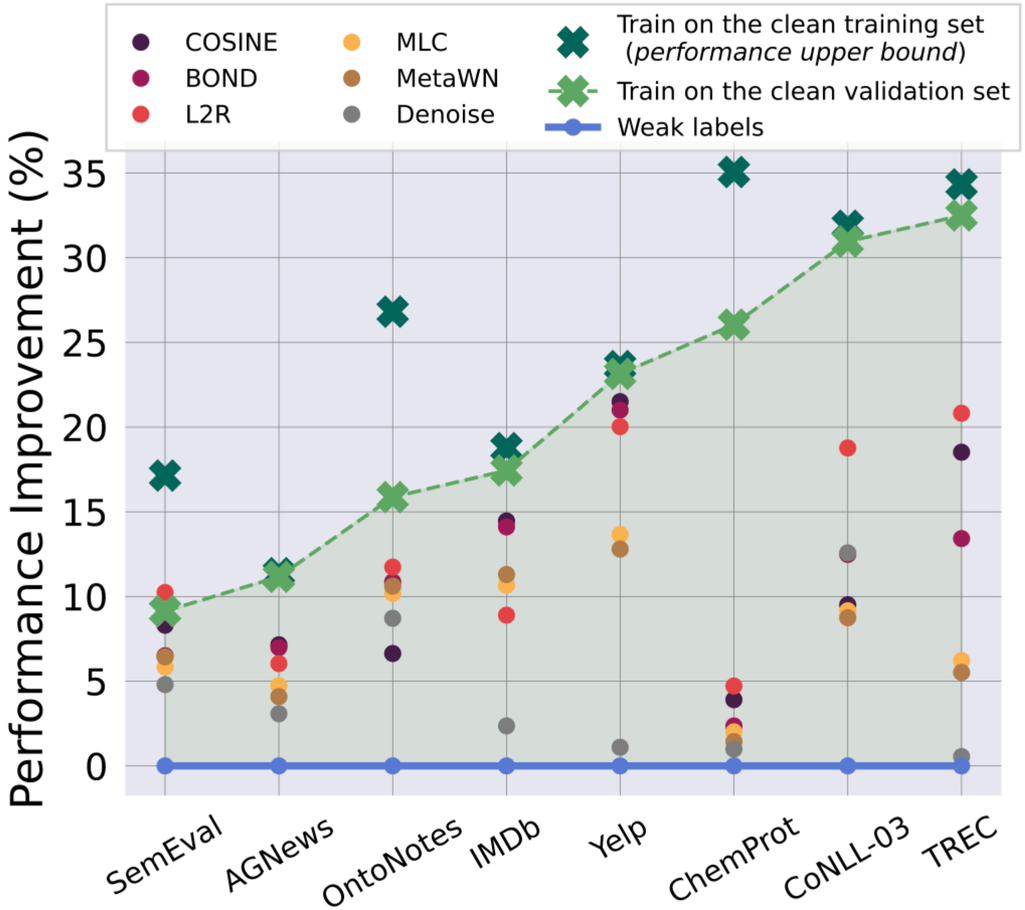
- ベースラインとして、弱い訓練データでモデルをfine-tuneした**FTw**も検証

実験１

- **弱教師あり学習と検証用データのみで学習を行なったモデルの性能を比較**
- WRENCHの各タスクについて弱教師あり学習と
  検証用データのみでの単純なfine-tuneをそれぞれ実施
- 評価用データにおいて生成則により決定されたラベルとの性能の差分を評価

実験１

- ほとんどのタスクにおいて
  弱教師あり学習は生成則によるラベルより
  性能が改善しているが、
  検証用データでfine-tuneした場合より
  性能が劣っている
  ↓
  クリーンなデータが少量存在する場合、
  弱教師あり学習よりも
  単純にfine-tuneした方が良い



実験２

- **クリーンな検証用データが無いとき、弱教師あり学習は機能するのか**

- WRENCHの各タスクについて弱教師あり学習を実施し、
  - クリーンな検証用データ
  - 生成則によりラベルを決定した弱い検証用データ
  - ランダム

  それぞれによりハイパラを選択し、性能を評価

- テストデータの弱いラベルによる性能をP_{WL}、

  弱教師あり学習の性能をP_{α}としたとき、

  $(P_{\alpha} - P_{WL})/P_{WL}$の値を性能の向上度合いとして定義

実験２

- クリーンな検証用データを用いた場合
  弱いラベルによる性能を上回っている
  （＝縦軸の値が0より大きい）

- 弱い検証用データを用いた場合
  弱いラベルによる性能とほぼ同等または低下し
  ランダムとほぼ変わらない結果

↓

クリーンな検証用データは不可欠であり、
これがない場合弱教師あり学習は機能しない

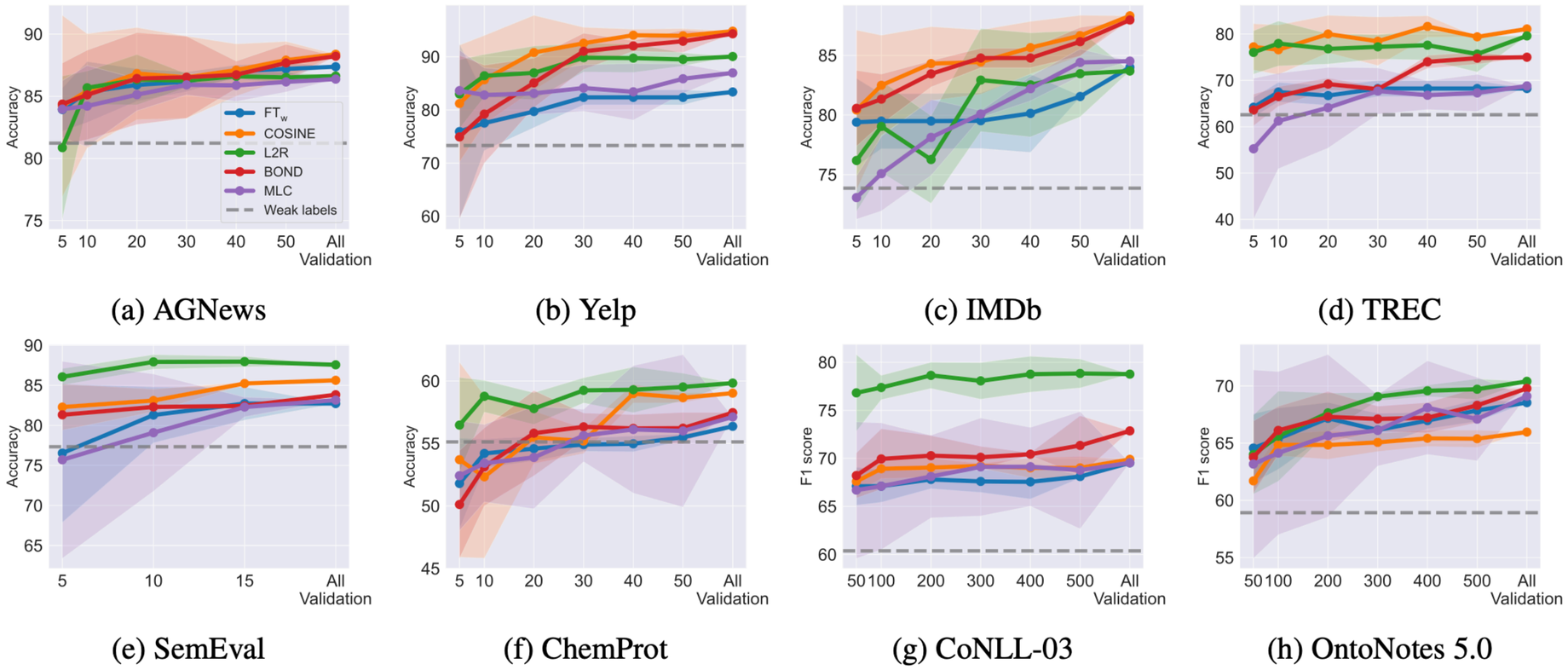
実験３

- **検証用データの数によって弱教師あり学習の性能はどう変化するか**

- 弱教師あり学習にはクリーンな検証用データが不可欠であるが、
  どのぐらいの量が必要なのかを検証

- WRENCHのうち固有表現認識タスク(**CoNLL-03,OntoNotes**)では
  {50, 100, 200, 300, 400, 500}、
  それ以外の分類タスクでは各クラス
  {5, 10, 15, 20, 30, 40, 50}の検証用データを用いたときの性能を評価

実験３

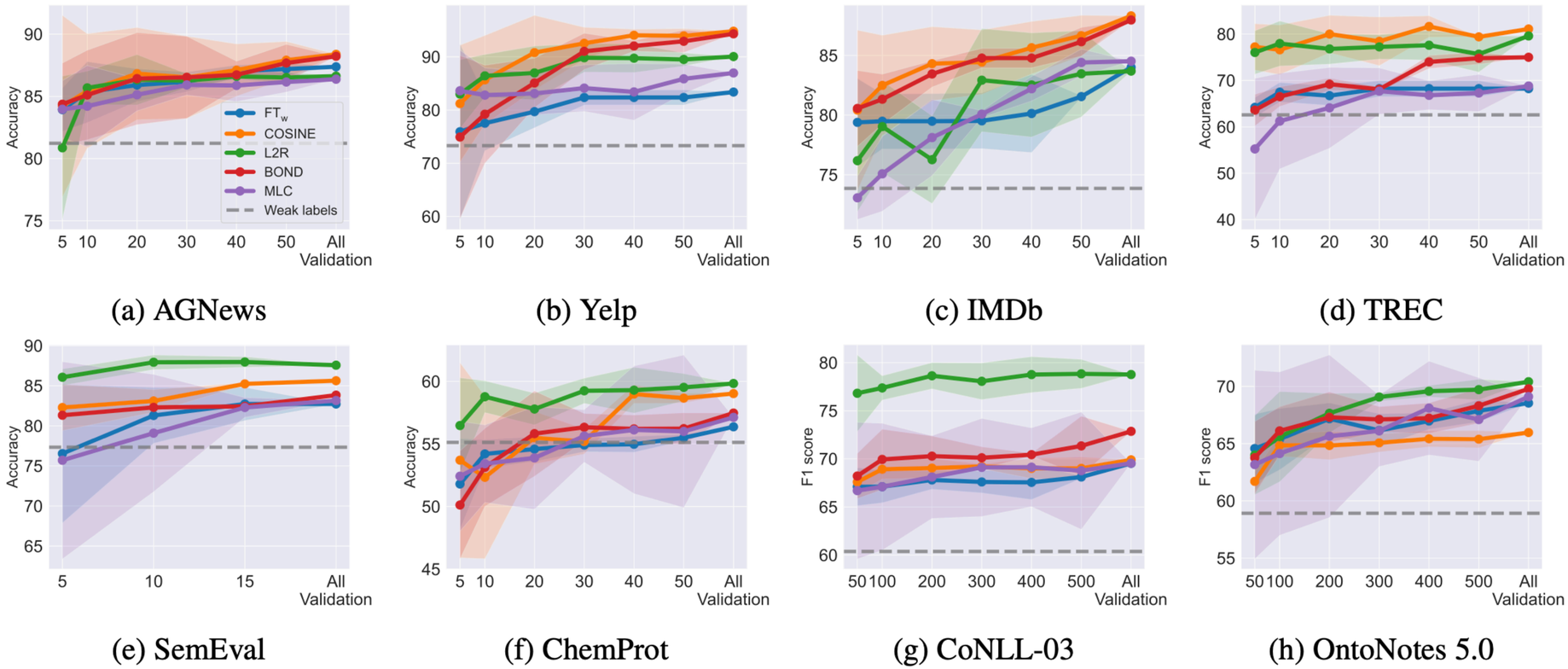


(a) AGNews (b) Yelp (c) IMDb (d) TREC

(e) SemEval (f) ChemProt (g) CoNLL-03 (h) OntoNotes 5.0

- 全体的にデータ数が増えると性能も向上する傾向があるが、30(a~f)/200(g,h)あたりから頭打ちになっている

実験３



(a) AGNews　(b) Yelp　(c) IMDb　(d) TREC

(e) SemEval　(f) ChemProt　(g) CoNLL-03　(h) OntoNotes 5.0

- 少ない検証用データ数でも弱いラベルのベースライン(灰色点線)を超えている
  ↓
  弱教師あり学習の有効性が示されている

実験３

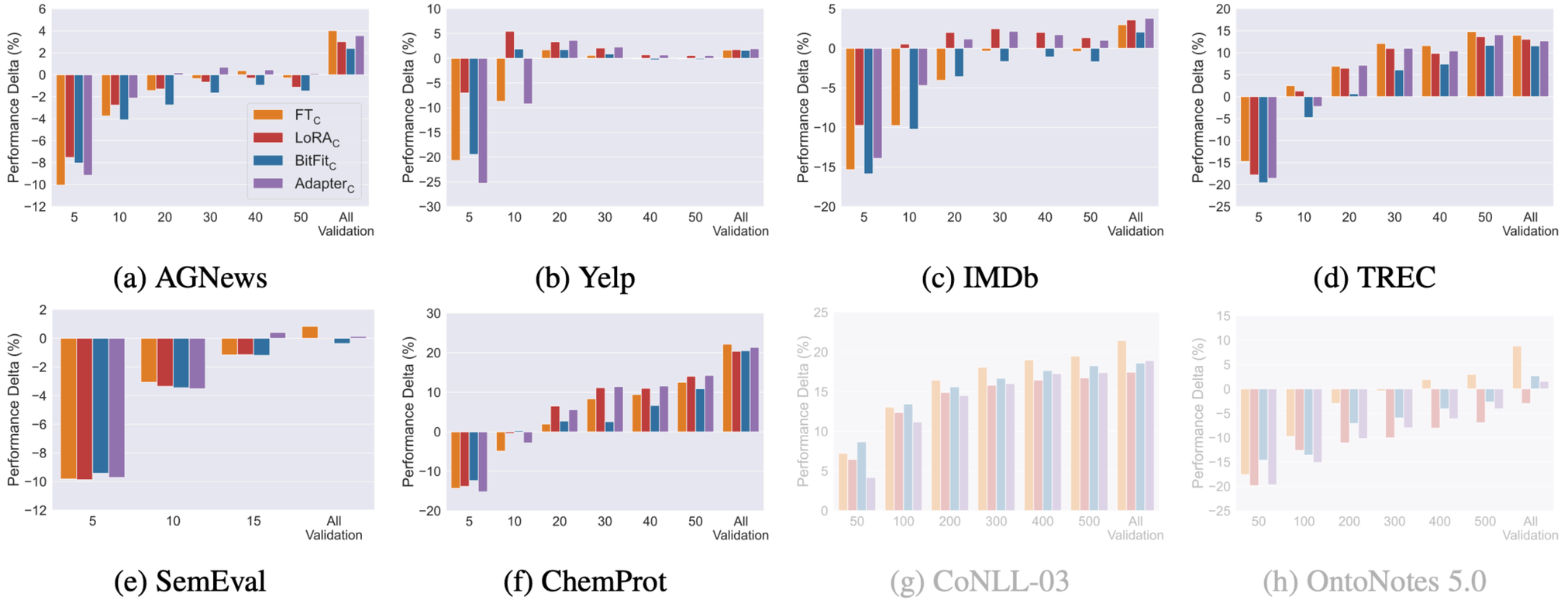
- 検証用データの数が少ないときでも弱教師あり学習は動作するが、
  検証用データでfine-tuneしたものと比較するとどうかを追加で検証

- 検証用データでfine-tuneを行うモデルとして、
  全てのパラメータの更新を行う**FT**に加え、
  LoRA、**BitFit**、**Adapter**を利用

- Fine-tuneを行うモデルでは6000ステップの学習終了時の
  モデルを評価に利用

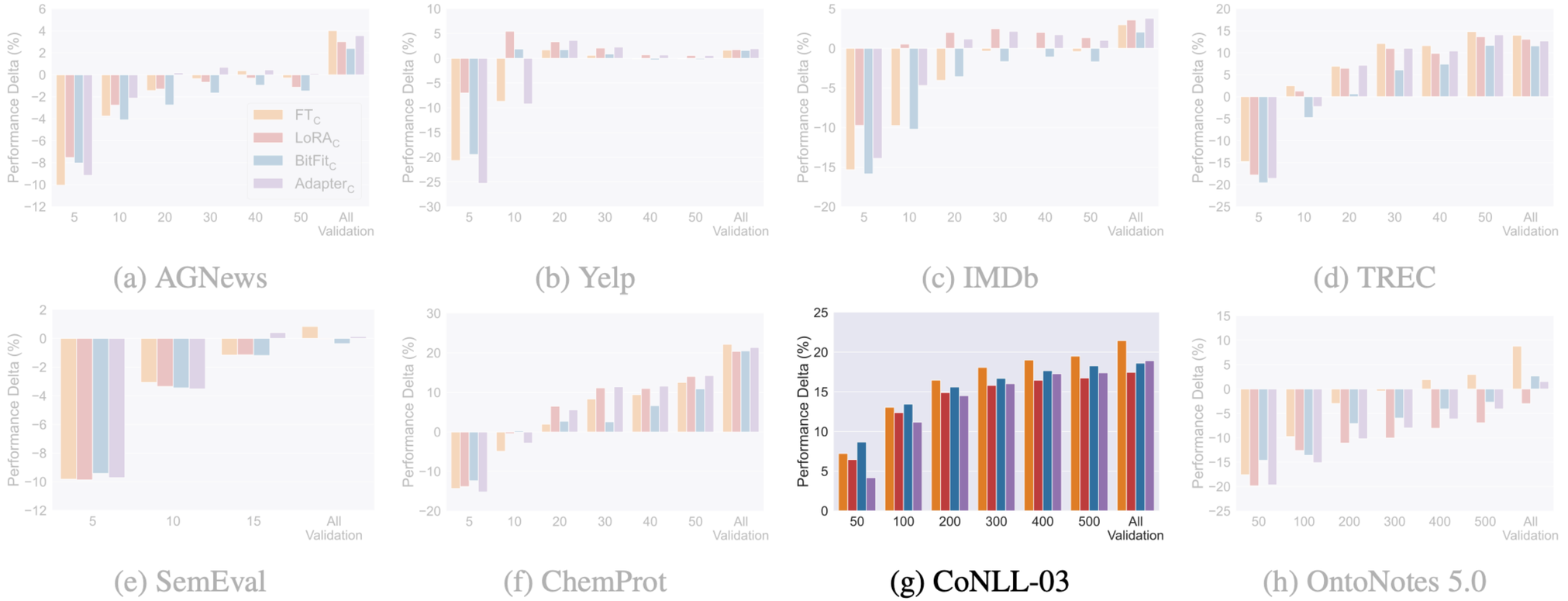
- 先ほどと同様の検証用データ数で実験し、
  弱教師あり学習のうち性能の良かった**COSINE**との性能を比較

実験３



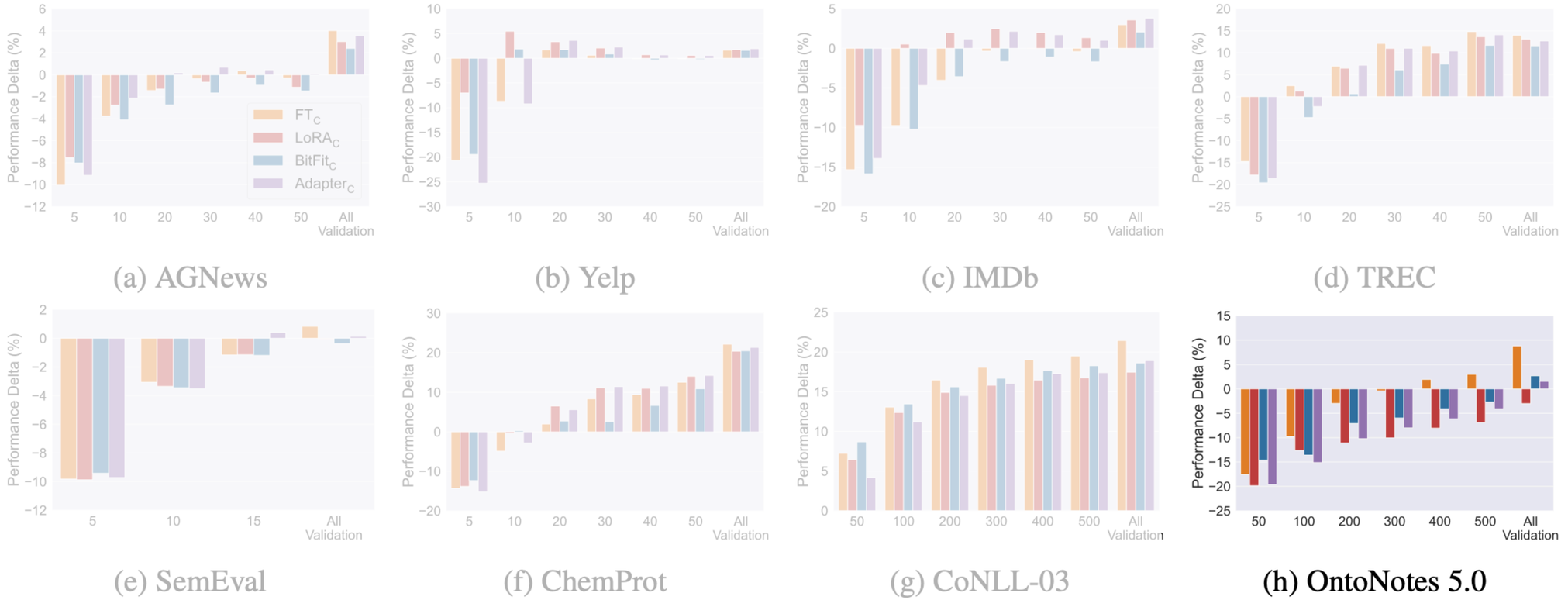
- 分類タスク(a~f)ではデータ数が10を超えたあたりから
  fine-tuneを行うモデルがCOSINEと同等、または上回っている
  （＝縦軸の値が0より大きい）

実験３



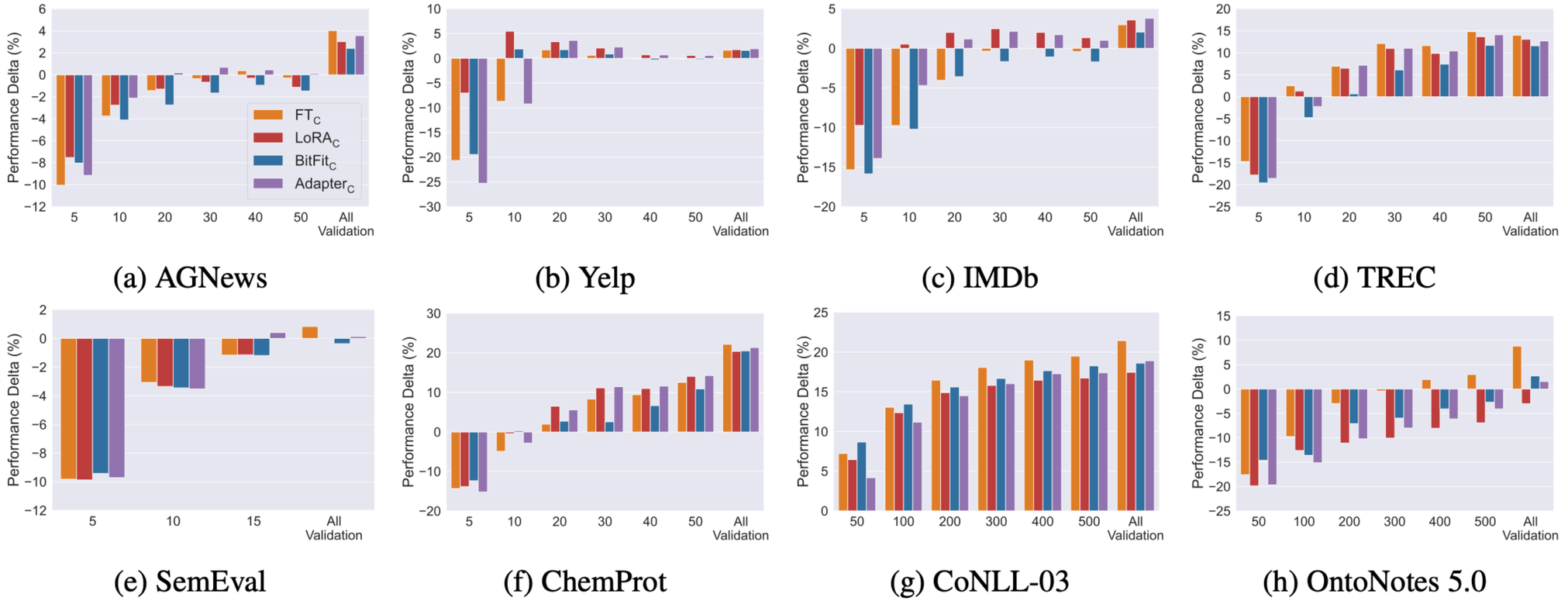
- CoNLL-03(固有表現認識タスク)では全ての手法が
  全てのデータ数の設定においてCOSINEを上回っている

実験３



- OntoNotes5.0(固有表現認識タスク)ではデータ数400で初めて
  COSINEを上回ることができたが、これでもCOSINEが学習に用いるデータ数の
  0.3%と非常に小さい

実験３



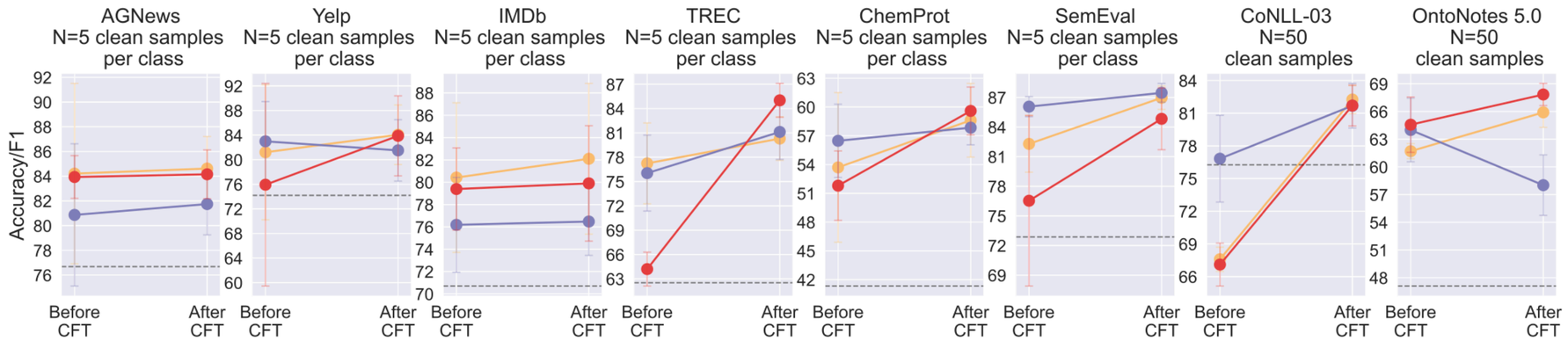
- 全体として、クリーンなデータが非常に少ない状況でのみ
  弱教師あり学習は実質的に有効であると示された

実験４

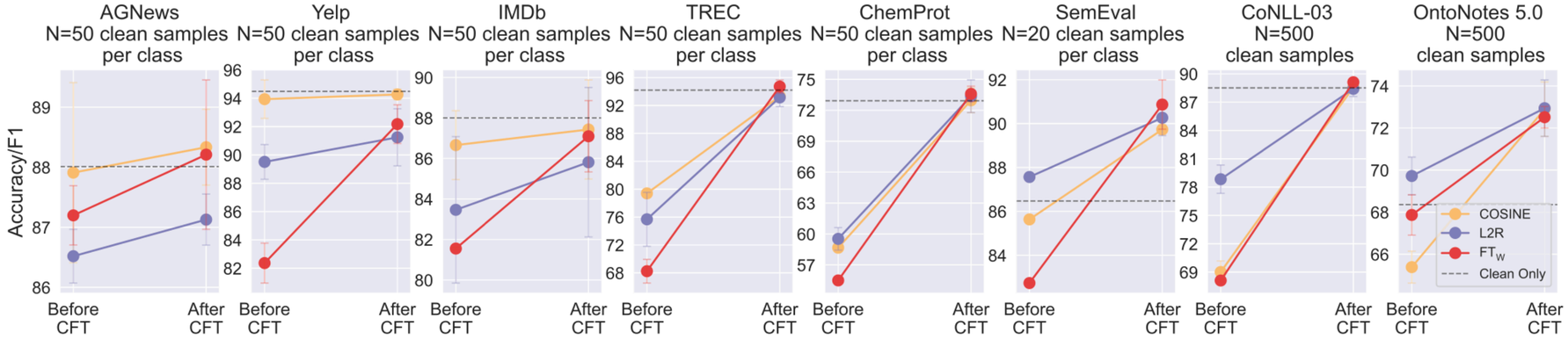
- **弱教師あり学習にてクリーンなデータで追加学習した場合性能はどう変化するか**
- 従来通り弱い教師データにて学習を行なったのち、クリーンな検証用データで追加で学習する**Continuous Fine-Tuning(CFT)**を導入
- **COSINE**、**L2R**、**FTw**にてCFTの有無による性能の差を評価
- 検証用データでの学習は6000ステップの学習終了時点でのモデルを評価に用いる

実験４

クリーンなデータ
5個



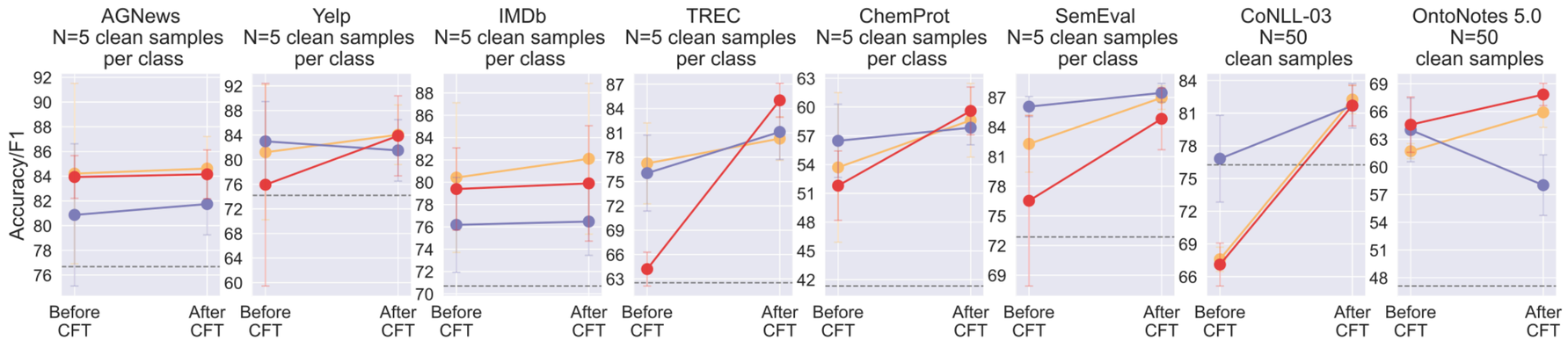
クリーンなデータ
50個



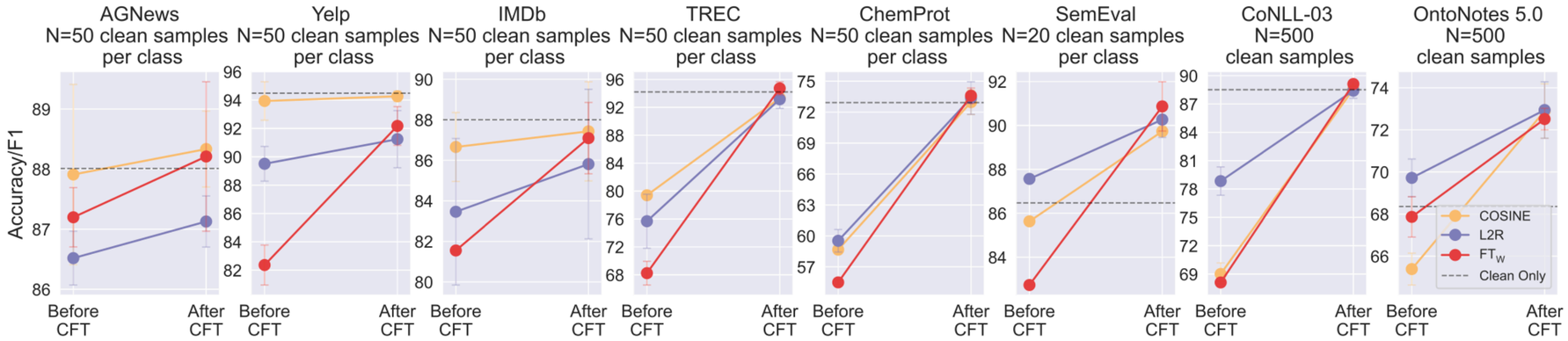
- 少ないデータ数でもCFTによって性能が改善するが、
  データが多いほど改善の幅は大きい傾向が見られた

実験４

クリーンなデータ
5個



クリーンなデータ
50個



- CFTによって**COSINE**、**L2R**と単純な手法である**FTw**の性能が近づいている
  ↓
  少量でもクリーンなデータが存在する場合
  複雑な弱教師あり学習の手法が非実用的であることを示している

実験４

- 弱教師あり学習とCFTの組み合わせが効果的となる要因を探るため、
  以下の二つの仮説を立てたのち、それぞれの検証のための実験を実施

  １.事前学習済み言語モデルが持つ知識が
  弱い教師データによるバイアスに抵抗しているのではないか

  ２.クリーンなデータの持つ情報が教師データのバイアスを軽減するのではないか

実験４ - 仮説１

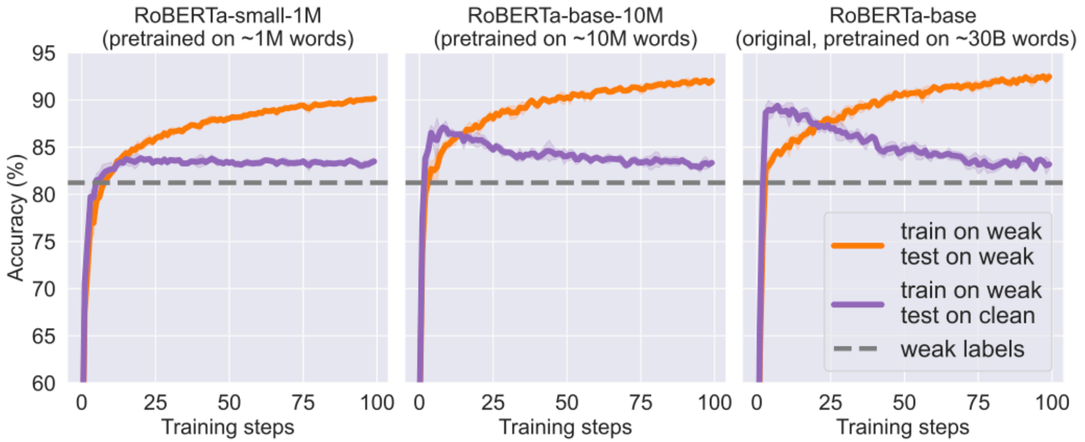
- **事前学習済み言語モデルが持つ知識が**
  弱い教師データによるバイアスに抵抗しているのではないか

- パラメータ数や事前学習のデータ数を変えた以下の3つのモデルを
  ベースとして利用し、性能の違いを検証
    - RoBERTa-small-1M
    - RoBERTa-base-10M
    - RoBERTa-base-30M

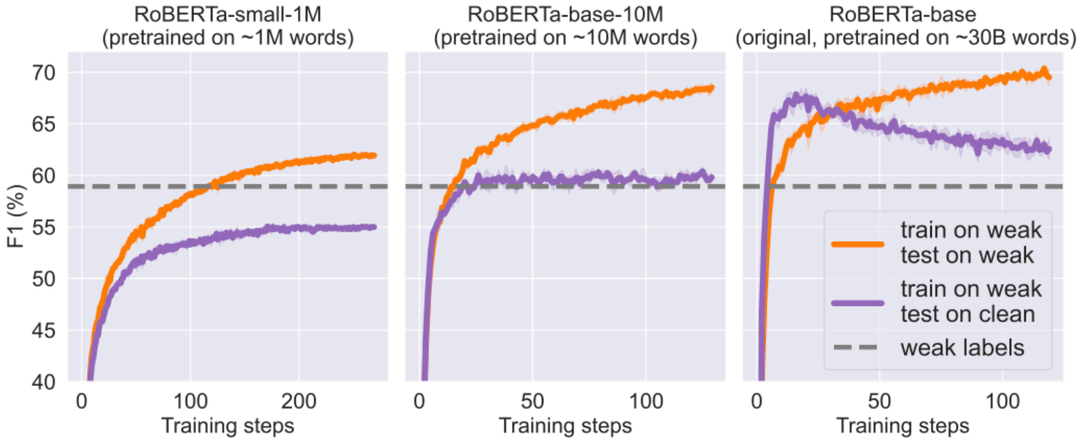
- 弱い教師データで学習を進めていったとき、
  クリーンなテストデータと弱いテストデータでの性能の推移を観察

実験４ - 仮説１

- 設定によらず学習が進むことで
  弱い教師データに適合していくが、
  （＝弱いデータでの性能が高く、
  クリーンなデータでの性能が低い）
  パラメータ数や事前学習のデータ数が
  大きいほど全体的な性能が高い



(a) AGNews

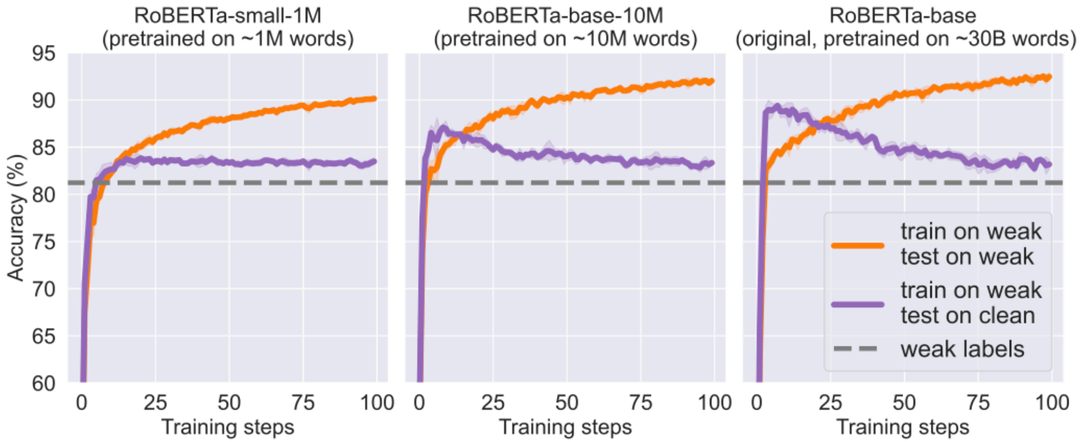


(b) OntoNotes 5.0

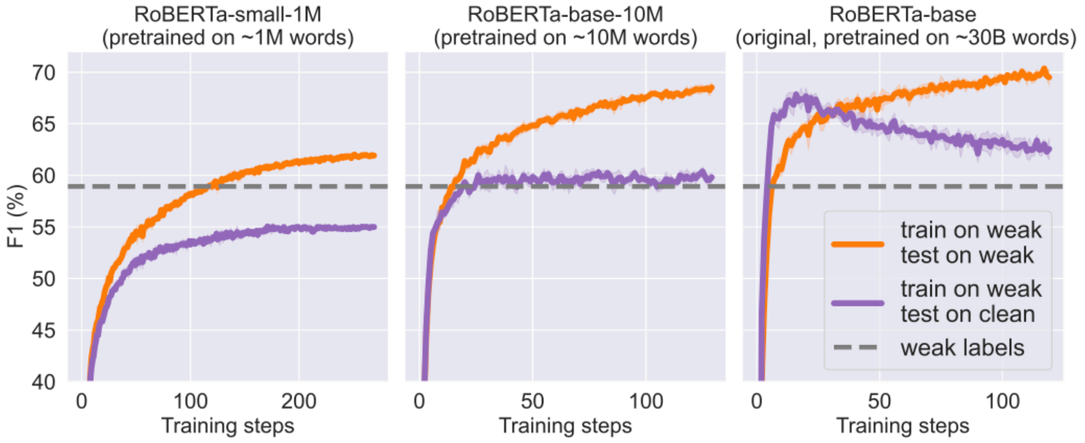
実験４ - 仮説１

- 事前学習のデータ数が大きいとき、
  学習の初期段階でクリーンなデータでの
  性能が顕著に高い
  ↓
  事前学習により、言語モデルは
  教師データから表層的な情報だけでなく
  より一般的な言語的関連を
  学習できる能力を得ている

- これにより単純な**FTw**でも
  学習に成功しているのでは？



(a) AGNews



(b) OntoNotes 5.0

実験４ - 仮説２

- **クリーンなデータの持つ情報が教師データのバイアスを軽減するのではないか**

- CFTでのクリーンなデータによる学習の際の
  クリーンなデータの**一致度合い**を変化させて性能を観察

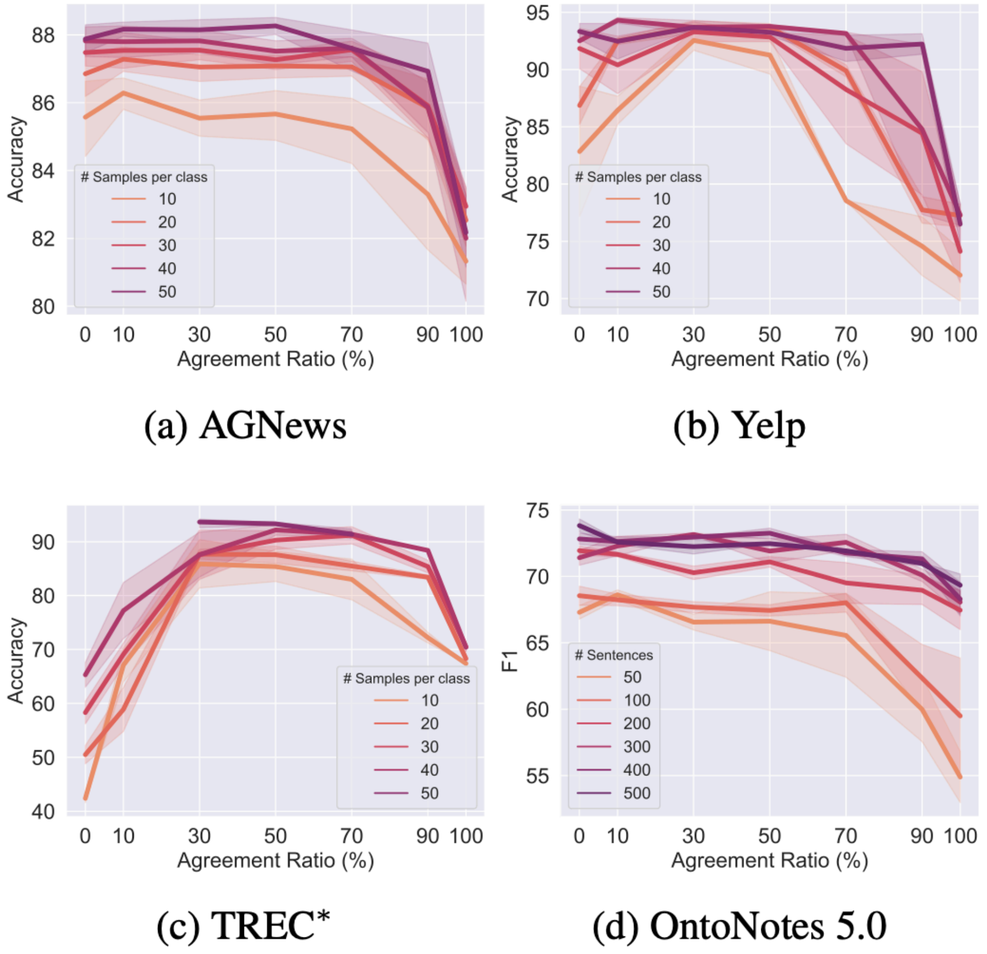
生成則による弱いラベルと
一致しているデータの割合

- 一致度合いが高いほど弱い教師データと差がなく、
  性能向上につながらないと考えられる

実験４ - 仮説２

- 全てのタスクにおいて
  一致度合いが70%を超えると性能が低下
  ↓
  弱い教師データの持つバイアスが
  保持されたままとなる

- 一致度合いが50%程度が最適であるが
  0%でもTREC以外ではある程度の
  性能が得られている
  ↓
  CFTにおいて、弱い教師データと矛盾する
  クリーンなデータがある程度必要なことが
  示された



まとめ

各実験より以下の知見が得られた

- 弱教師あり学習にて検証用にクリーンなデータを用いる場合、
  その検証用データでfine-tuneした方が性能が良いことが多い

- クリーンなデータが無いと弱教師あり学習は機能しない

- クリーンなデータが少しでもあれば弱教師あり学習はある程度機能し、
  クリーンなデータでfine-tuneしたモデルより性能が上回る

- 弱い教師データの後検証用のクリーンなデータでも学習を行う
  CFTが弱教師あり学習に対して有効である

まとめ

これ以降弱教師あり学習手法を提案するときは以下が推奨される

- 提案手法のモデル選択基準、特にクリーンなデータの存在に
  どの程度依存しているかを明示する

- 提案手法に匹敵する性能を単純なfine-tuneで達成するために
  どれぐらいクリーンなデータが必要かを明記する
  （少量のみ必要な場合、弱教師あり学習が最適でない可能性がある）

- 検証用などにクリーンなデータが必要な場合、
  CFTを実施した性能を報告する