単一評価サンプルのためのトランズダクティブ学習

佐々木翔大 [1,2]　大内啓樹 [1,2]　鈴木潤 [2,1]　Ana Brassard[1,2]　乾健太郎 [2,1]
[1] 理化学研究所　[2] 東北大学
{shota.sasaki.yv, hiroki.ouchi, ana.brassard}@riken.jp
{jun.suzuki, inui}@ecei.tohoku.ac.jp

1 はじめに

トランズダクティブ学習の目的は，未知データに対するモデルの汎化性能を向上させることではなく，手元にある限られた評価データに対する性能を向上させることである．この設定において，評価データはラベル無し訓練データとして，ラベル有り訓練データと合わせて活用することができる．トランズダクティブ学習は昨今，新しいタスク設定として自然言語処理（NLP）コミュニティにおいて注目を集めている．例えば，大内ら [9] は評価データに特化した単語分散表現を獲得するために，ラベル無しの評価データにおいて言語モデルの学習を追加で行う手法を提案した．Poncelas ら [11] は機械翻訳のタスクにおいて，評価データとの類似性に基づきサンプリングした訓練データの部分集合を用いて，モデルを再学習する手法を提案した．両手法とも，評価データが与えられた後に追加で学習する十分な時間が確保できる場合，トランズダクティブ学習を用いることで性能を向上が見込めることを実験的に示した．

本研究では，より限定的な状況におけるトランズダクティブ学習の応用可能性と限界点を調査する．一般的なNLP タスクにおいて，評価データはある一定量（1000 サンプルなど）が確保されている場合がほとんどである．それゆえ，既存の手法は暗黙的に一定量の評価データを使用可能な状況を仮定していると言える．しかしながら現実世界においては，非常に限られた評価データしか手に入らない状況も十分考え得る．例えば，Web で提供されるエッセイスコアリングサービスを考えると，評価データは不確定なタイミングで一つずつ提出されるだろう．そこで本研究では「評価データのサンプル数が非常に限られた状況でトランズダクティブ学習を行う場合，依然として性能の向上を確認できるか」という問いに関して調査を行う．具体的には，最も極端な状況，つまり評価データのサンプル数が 1 つであるという設定におけるトランズダクティブ学習（図 1）の効果を，固有表現認識，感情分類，含意関係認識の三つのタスクを用いて検証する．

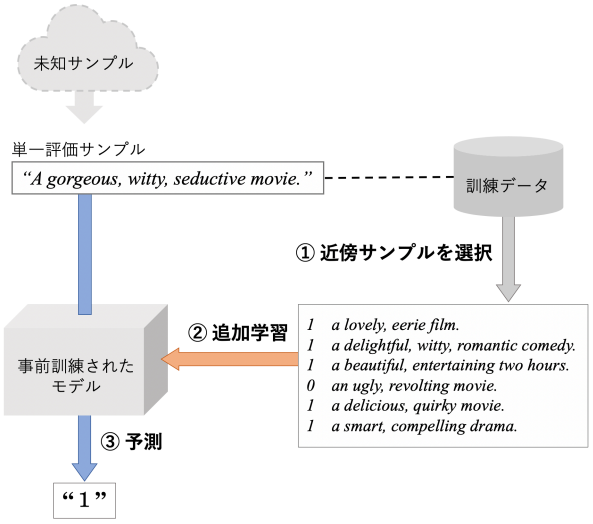

図1: 映画レビューの感情分類における単一評価サンプルのためのトランズダクティブ学習

2 関連研究

トランズダクティブ学習は Vapnik [15], Gammerman ら [4] によって提唱された．以降，トランズダクティブ学習手法は NLP 外で広く研究されている [1, 7, 12]．NLP におけるトランズダクティブ学習に関して言えば，Joachims[6] と Ifrim ら [5] がテキスト分類タスクにトランズダクティブ学習を適用した．最も最近の研究としては，大内ら [9] の研究が挙げられる．彼らは，言語モデルを大規模なラベル無しコーパスで事前学習した後，さらにラベル無し評価データも用いて追加で学習を行っている．学習された言語モデルを評価データに特化した埋め込み層として用いて，タスク毎のモデルの訓練を行った．Poncelas ら [11] は機械翻訳タスクにトランズダクティブ学習の適用することを試みた．しかしながら，既存のトランズダクティブ学習に関する研究においては，ある一定量の評価データを前提としており，より現実問

Algorithm 1 単一サンプルトランズダクティブ学習

入力: $\mathcal{D}^{\text{train}} = \{(X_i^{\text{train}}, Y_i^{\text{train}})\}_{i=1}^{N}$,
$\mathcal{D}^{\text{test}} = \{X_i^{\text{test}}\}_{i=1}^{M}$,

1: $\Theta' \leftarrow \arg\min_{\Theta} L(\Theta|\mathcal{D}^{\text{train}})$
2:
3: **for** $i = 1 \ldots M$ **do**
4: $\quad \mathcal{D}_i^{\text{train-sub}} \leftarrow \text{Sample}(\mathcal{D}_i^{\text{test}}, \mathcal{D}^{\text{train}})$
5: $\quad \Theta_i'' \leftarrow \arg\min_{\Theta'} L(\Theta'|\mathcal{D}_i^{\text{train-sub}})$
6: $\quad \text{Outputs} \leftarrow \text{Predict}(\mathcal{D}_i^{\text{test}}|\Theta_i'')$

題に即した単一評価サンプルを前提とした設定における試みはなされていない．

3 タスク設定

はじめに，大内ら [9] に基づき，トランズダクティブ学習の定義を記述する．訓練データを $\mathcal{D}^{\text{train}} = \{(X_i^{\text{train}}, Y_i^{\text{train}})\}_{i=1}^{N}$，評価データを $\mathcal{D}^{\text{test}} = \{X_j^{\text{test}}\}_{j=1}^{M}$ とした時，トランズダクティブ学習においては，与えられた評価データ $\mathcal{D}^{\text{test}}$ に対して，即時に応答する必要はないことが前提である．より具体的には，$\mathcal{D}^{\text{train}}$ と $\mathcal{D}^{\text{test}}$ を用いてモデルを再学習する十分な時間が与えられているということである．ここで，システムが $\mathcal{D}^{\text{test}}$ を受け取ってから出力を返さなければならない最小時間を T とし，$\mathcal{D}^{\text{train}}$ と $\mathcal{D}^{\text{test}}$ を用いてモデルを学習するのに必要な最大時間を T^{train} とすると，トランズダクティブ学習は以下のように説明できる．

定義 (Transductive 学習). トランズダクティブ学習の目的は，$T^{\text{train}} < T$ の条件を満たしながら，$\mathcal{D}^{\text{test}}$ に対する予測の正解率を最大化することである．

また，本研究では推論時において以下のような仮定を置く．

仮定 (単一評価サンプル) 評価データは単一データサンプルからなる（$M = 1$）．

4 提案手法

本節では，提案手法である単一評価サンプルのためのトランズダクティブ学習手法について説明する．Algorithm 1は提案手法の擬似コードである．まずはじめに，$\mathcal{D}^{\text{train}}$ を用いてモデルを事前訓練する（行 1）．次に与えられた評価サンプル $\mathcal{D}_i^{\text{test}}$ と訓練データ $\mathcal{D}^{\text{train}}$ 内の全てのデータとの類似度を測る（行 4）．例えば入力がテキストの場合は，入力を事前学習された埋め込みベクトルで表現し，ベクトル同士の類似度を測ることなどが考えられる．次に算出した類似度の基づき，評価サンプルに対する K 個の近傍サンプルを訓練データ $\mathcal{D}^{\text{train}}$ から選択し（行 4），それを用いてモデルの追加学習を行う（行 5）．例えばテキスト分類の場合で言えば，モデルのパラメータは一時的に類似した文集合を用いてチューニングされる．この手順は仮に複数の評価サンプルがある場合でも，それぞれの追加学習に依存関係はないため，全ての新しい評価サンプルに対して並行して処理することができる．

この手法は，トランズダクティブ学習の設定において，モデルは未知のデータに対する汎化性能を有する必要はなく，ただ単一の評価サンプルに対して特化すればよいという考えに基づいている．重要なことには，提案手法は評価サンプル数に依存しない手続きからなるため，評価サンプルがいくつある場合でも適応可能であるということである．つまり提案手法は，手に入る評価サンプルの量が不確定であるという現実世界の状況に，より柔軟に対応できる手法であるといえる．一方，既存のトランズダクティブ学習の手法は，十分な数の評価サンプルが手元に揃っていることを前提としている．

5 実験

本節では，単一評価サンプルを前提とした我々の手法を，トランズダクティブ学習を用いない手法と，一定量の評価データを用いたトランズダクティブ学習手法の 2 つの手法と比較する．

5.1 タスク

実験に用いるタスクは以下の 3 つである．

- 固有表現認識（NER）タスク（CoNLL-2003 データセット [14]）
- 感情分類タスク（Stanford Sentiment Treebank (SST-2) データセット [13]）
- 含意関係認識タスク（RTE データセット [2]）

5.2 実験設定

提案手法で近傍サンプルを選択する際の入力サンプルの基本単位をどう定めるかは自明ではない．本研究では，NER, SST-2 の実験においては，1 文を 1 サンプルとし，RTE の実験においては，テキスト文と仮説文を連結したものを 1 サンプルとした．サンプル間の類似度の指標としては，Zyang ら [16] によって提案された BERTScore を用いた．BERTScore は BERT[3] の埋め込みに基づき，意味的及び文法的な観点から 2 文間の類

表1: NER，SST-2，RTE の実験結果

手法	NER	SST-2	RTE
BASE	91.9	96.1	77.3
FINE-TUNED FULL	92.1	96.3	**79.4**
FINE-TUNED SINGLE	**92.3**	**96.4**	79.1

似度を計算する手法である．

NER の実験では，単語ベクトル層，CNN を用いた文字ベクトル層，2 層の Bi-LSTM 層，CRF 層から構成される．単語ベクトル層には ELMo [10] を用いた．また，各評価サンプルに対する近傍サンプルの数は $K = 500$ とした．テキスト分類問題である SST-2, RTE の実験では，BERT [3] の構造を基にした RoBERTa [8] を用いた．テキスト分類において RoBERTa を用いる際は，入力テキストを分散表現にエンコードした後，線形変換と softmax 関数を用いて分類ラベルを予測するという Liu ら [8] の設定に従った．各評価サンプルに対する近傍サンプルの数は $K = 1000$ とした．

評価時には他の手法と公平に比較するため，各評価サンプル X_j^{test} に対して提案手法で 1 つずつ予測した後，標準的な評価手順で性能評価を行った．また，SST-2, RTE の実験は評価データの正解ラベルが公開されていないため，開発データにおける性能を報告する．

5.3 比較対象

以降，単一評価サンプルの近傍サンプルでモデルを追加で学習する提案手法を FINE-TUNED SINGLE と呼称する．これに対する一つ目の比較対象として，トランズダクティブ学習を用いない手法，つまり各データセットの訓練データのみ用いて学習する手法を BASE と呼称する．次に，二つ目の比較対象として，用意された評価データの全てを訓練に用いるという典型的なトランズダクティブ学習手法を FINE-TUNED FULL と呼称する．具体的には．各評価サンプルに対して選択された $M \times K$ 個の近傍サンプル（重複を含む）を選択回数順に並べ，選択回数の高い方から K 件の近傍サンプルを用いて追加学習を行った．

5.4 実験結果

NER の実験における F1 スコア，SST-2，RTE の実験における正解率を表 1に示す．全てのタスクにおいて，FINE-TUNED SINGLE と FINE-TUNED FULL は BASE の性能を上回った（+0.2 ～ +2.1 ポイント）．このことから，これらのタスクにおいてトランズダクティブ学習が有効であることがわかった．また，FINE-TUNE SINGLE の性能は FINE-TUNE FULL を上回る，もしくは同等の性能であった．このことから，提案手法は単一評価サンプルの情報を利用することで，多くの評価サンプルの分布情報を利用できるような状況ではない限られた状況においても性能の向上を達成できていると言える．

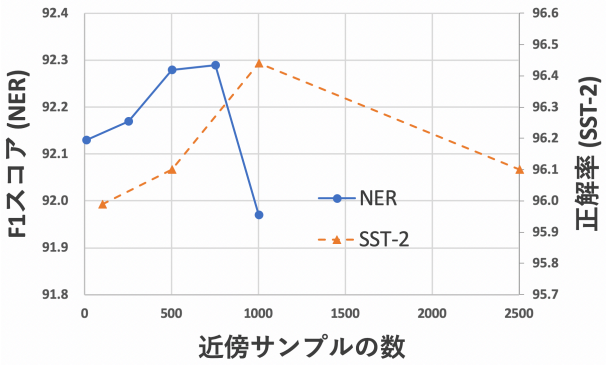

図2: 近傍サンプルの数と性能の関係

表2: BASE と FINE-TUNED SINGLE の予測の変化

	FINE-TUNED T	FINE-TUNED F
BASE T	45592 (38070)	46 (22)
BASE F	71 (11)	726 (220)

6 分析

本節では (i) 近傍サンプルの数と性能の関係, (ii) BASE によって間違って予測されたが，FINE-TUNED SINGLE によって正しく予測された評価サンプルについて，分析を行った．

6.1 近傍サンプルの数と性能の関係

図 2は近傍サンプルの数と性能の関係を表すグラフである．青色の実線は NER，橙色の破線は SST-2 における性能を表している．概して，近傍サンプルの数が 500 ～ 1000 個の範囲において，最も良い性能に達することがわかった．しかしながら，訓練データのカバレッジはタスク間で大きく異なるため，最適な近傍の数はタスク依存であると言える．つまり，訓練データのカバレッジが低いにも関わらず，必要以上に多くの近傍サンプルを利用し過ぎても，評価サンプルの予測に悪影響を及ぼしかねない．一方で，近傍サンプルの数が少な過ぎても，追加の学習に十分なサンプル数を確保できず，結果的に性能向上が見込めない場合も考えられる．

6.2 予測の変化

表 2は NER の実験における予測の変化を表す表である．‘T’ は固有表現タグを正しく予測したこと，‘F’ は固有表現タグを誤って予測したことを表す．例えば，表中

		太字の単語のタグ
Test	Leeds had already fined **Bowyer** 4,000 pounds ($6,600) and warned him a repeat of his criminal behaviour could cost him his place in the side.	B-ORG → B-PER
NN 1	**Cyprien**, also fined 10,000 Swiss francs ($8,400), traded punches with St Gallen's Brazilian player Claudio Moura after a match on Saturday.	B-PER
NN 2	**Moura**, who appeared to have elbowed Cyprien in the final minutes of (...), was suspended for seven matches and fined 1,000 francs ($840) (...).	B-PER
NN 3	Romania's soccer bosses also fined **Cozma**, a well-known miners' union leader, 10 million lei ($3000) for the half-time attack on Dinamo Bucharest's (..).	B-PER

表3: NER の実験における評価サンプル（Test）とそれに対して提案手法によって選択された近傍サンプル（NN）の例

の左上の値は BASE と FINETUNED-SINGLE の両方によって正しく予測された固有表現タグの数を表している．括弧内の値は，固有表現でないことを表す 'O' タグの数を表している．BASE が予測を誤ったが，FINETUNED-SINGLE が正しく予測したタグの数は 71 個で，その逆は 46 個であった．手作業で 46 個のサンプルを確認したところ，そのうち 28% が正解ラベル自体が誤って付与されており，結果的に予測が誤りであると判断されてしまったサンプルであった．

表3は BASE が予測を誤ったが，FINETUNED-SINGLE が正しく予測した評価サンプルに対して，その時選択されていた近傍サンプルの例である．BASE は人名である *Bowyer* に対して組織を表す B-ORG タグを誤って予測してしまっていたが，FINETUNED-SINGLE は人名を表す B-PER タグを正しく予測した．FINETUNED-SINGLE においては，近傍サンプル内の人名に付与された B-PER タグを手がかりに正しい予測をするよう学習できたと考えられる．

7 おわりに

本研究では，利用できる評価サンプルの数が 1 つに限られる設定におけるトランズダクティブ学習手法を提案し，その有効性を検証した．評価サンプルの数が非常に限られている状況においても，単一の評価サンプルの近傍サンプルを用いてモデルを追加で学習することで，性能の向上が見込めることがわかった．

謝辞 本研究は JSPS 科研費 JP19K20351, JP19H04162 の支援を受けて行った．

参考文献

[1] Andrei Alexandrescu and Katrin Kirchhoff. "Graph-based learning for statistical machine translation". In: *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. 2009, pp. 119–127.

[2] Luisa Bentivogli et al. "The Fifth PASCAL Recognizing Textual Entailment Challenge". In: *Proceedings Text Analysis Conference (TAC)*. 2009.

[3] Jacob Devlin et al. "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 (NAACL)*. 2019, pp. 4171–4186.

[4] A. Gammerman, V. Vovk, and V. Vapnik. "Learning by Transduction". In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Madison, Wisconsin, 1998, pp. 148–155.

[5] Georgiana Ifrim and Gerhard Weikum. "Transductive Learning for Text Classification Using Explicit Knowledge Models". In: *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*. 2006, pp. 223–234.

[6] Thorsten Joachims. "Transductive Inference for Text Classification Using Support Vector Machines". In: *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*. 1999, pp. 200–209.

[7] Yanbin Liu et al. "Learning to Propagate Labels: Transductive Propagation Network for Few-Shot Learning". In: *International Conference on Learning Representations (ICLR)*. 2019.

[8] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[9] Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. "Transductive Learning of Neural Language Models for Syntactic and Semantic Analysis". 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)*. Nov. 2019, pp. 3656–3662.

[10] Matthew Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. June 2018, pp. 2227–2237.

[11] Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. "Transductive Data-Selection Algorithms for Fine-Tuning Neural Machine Translation". In: *Proceedings of The 8th Workshop on Patent and Scientific Literature Translation*. 20 8 2019, pp. 13–23.

[12] Ozan Sener et al. "Learning Transferrable Representations for Unsupervised Domain Adaptation". In: *Advances in Neural Information Processing Systems 29 (NIPS)*. 2016, pp. 2110–2118.

[13] Richard Socher et al. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2013, pp. 1631–1642.

[14] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*. 2003, pp. 142–147.

[15] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[16] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *arXiv preprint arXiv:1904.09675* (2019).