

低リソースの意見要約のための教師なしドメイン適応

小杉宥斗¹ 李吉屹² 福本文代²

¹ 山梨大学工学部コンピュータ理工学科 ² 山梨大学大学院総合研究部工学域
{t18cs021, jyli, fukumoto}@yamanashi.ac.jp

概要

近年、意見要約の分野では、Few-Shot 学習を利用した研究が盛んに行われている [1][2][3]。しかし、既存の Few-Shot 学習モデルでは特定のドメインの教師データが存在しない場合、そのドメインにおける要約の精度が低下する。この問題の解決を図るため、本研究では MoE (Mixture of Experts) による教師なしドメイン適応に基づく新たな手法を提案する。作成したモデルに対する評価実験とその結果から得られた課題について報告する。

1 はじめに

意見要約 (Opinion Summarization) とは主観的な意見を含む複数のテキストを対象とし、それらの意見を端的にまとめたテキストを生成するタスクである。本研究では、製品に対するユーザーレビューを対象とし、深層学習を用いた意見要約を行う。教師あり学習により意見要約モデルを訓練する場合、大量の教師データを用意する必要があり、多大なコストや労力を要する。そのため、意見要約分野では教師データが不足しており、従来より教師なし学習や半教師あり学習を用いた手法が数多く提案されている [4][5][6][7][8][9]。特に近年では、Few-Shot 学習を用いた手法が盛んに研究されている [1][2][3]。Few-Shot 学習とは、ごく少数の教師データのみを利用してモデルを訓練する学習手法である。

深層学習では、しばしば特定のドメインにおける教師データが存在しない場合が生じる。これを解決するための一手法として、教師なしドメイン適応 (Unsupervised Domain Adaptation) がある。教師なしドメイン適応は、教師データが利用できる「ソースドメイン」によりモデルを訓練し、教師データが存在しない「ターゲットドメイン」における汎化性能を獲得する技術である。特に、ユーザーレビューの対象となる製品は様々なドメインに属している。これらの知識を有効に活用するため、本研究

では複数のソースドメインを利用したドメイン適応 (Multi-Source Domain Adaptation) に焦点をあてる。近年では、このようなドメイン適応に対し MoE を利用した手法が提案され、その有効性が示されている [10][11][12]。

本研究では、Few-Shot 学習を利用した意見要約モデルである FewSum[1] に対し、MoE (Mixture of Experts) による教師なしドメイン適応を行うモデルとして、FewMoESum を提案する。加えて、モデルに対する評価実験とその結果から得られた課題について報告する。

2 FewSum

FewSum[1] は Few-Shot 学習による意見要約の代表的な手法であり、入力レビュー集合に出現しない表現も用いて要約を作成する Abstractive Summarization を行う。モデルは Transformer[13] を採用しており、Encoder と Generator (Decoder) から構成される。

FewSum の学習は、事前学習と Fine-tuning からなる。事前学習は、複数のレビューをもとにレビューを生成することにより行う。学習時は、同じ製品に対するレビューの集合から 1 つを抜き出したレビュー r_i を教師信号とし、残りのレビュー r_{-i} を入力サンプルとする。モデルのうち、Encoder は入力された r_{-i} の特徴量を抽出し、Generator は抽出された r_{-i} の特徴量と property をもとにレビュー r_i を推測する。property とは r_i と r_{-i} の関係や、 r_i の性質であり、事前学習時点では r_i と r_{-i} に応じて一意に計算される。事前学習により、Generator は property と生成すべきテキストの対応関係を学習する。

Fine-tuning では、人手で作成したごく少数の要約を教師信号として利用し、要約生成モデルを学習する。ここでは、Plug-in Network、すなわち、入力レビュー集合から抽出した情報をもとに、理想的な要約の生成を促す property を予測するよう学習するネットワークをモデルに追加する。

property の例には「代名詞の分布」がある。要約

は、一人称代名詞を含まない客観的文体であることが望ましい。そのため、Plug-in Network は「一人称代名詞の割合が小さい」ことを示す property を生成する。Generator は事前学習時に property に対応したテキストを生成するよう学習しているため、当該の property が与えられることにより、一人称代名詞の使用を避けつつ要約の生成を行うことができる。

Few-Shot 学習では教師データがごく少数に限られており、大量のパラメータを調整するには不十分であるため、モデル全体に対する Fine-tuning を行うと過学習が発生する。そのため Fine-tuning 時には、調整するパラメータを Plug-in Network と Generator の一部に限定することにより過学習を抑制している。

3 Mixture of Experts (MoE)

Guo らは、感情分類や品詞タグ付けといった分類タスクを対象とし、MoE (Mixture of Experts) を用いた教師なしドメイン適応を行う手法を提案している [11]。MoE は、特定のデータの処理に特化したネットワークを複数作成し、各ネットワークの出力を加重平均する手法である。特に、MoE ではこれらネットワークを「エキスパート」と呼んでいる。Guo らの手法 (以降、GuoMoE と記述する) では、複数の分類器をそれぞれ各ソースドメインに特化し、エキスパートとしている。図 1 にその概略図を示す。分類を行う際は、入力サンプルとソースドメインの類似度が高いほど、そのソースドメインに特化したエキスパートに対して大きい重みを割り当てる。この重みに基づき各エキスパートの出力を加重平均し、最終的な分類結果を求める。この処理により、入力サンプルに類似するソースドメインから得た知識を重視して分類を行うため、未知のドメインのデータに対しても優れた分類結果を得ることができる。

GuoMoE と FewSum との差異を以下に示す:

- (a) GuoMoE では、大量の教師データを用い教師あり学習を行っている。一方、FewSum ではごく少数の教師データを用いる Few-Shot 学習を行う。
- (b) GuoMoE では、実験に使用したモデルは小規模である。一方、FewSum では比較的大規模なモデルを扱う。したがって Generator をエキスパートとして単純に複製すると、モデルのパラメータ数が大幅に増加する。

大量のパラメータを調整する場合、それに応じて大量の教師データを利用しなければ過学習が発生す

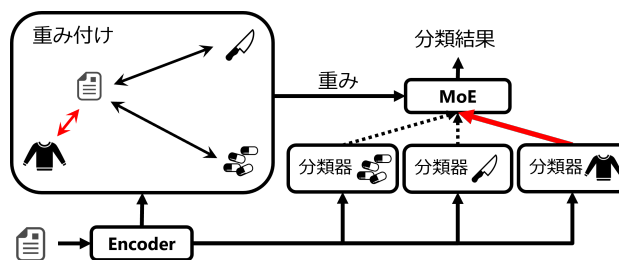


図 1 Guo らの手法 (GuoMoE)

る。そのため、GuoMoE のようにネットワークを単純に複製する方法は、FewSum のような条件のもとでは適さない。したがって本研究では、FewSum の Generator のうち一部のみをエキスパートとすることにより、この問題の解決を図る。

4 提案手法

本手法では、事前学習と FewMoESum の学習によりモデルを訓練する。事前学習の際は、FewSum と共通のモデル、及び学習方法を用いる。FewMoESum の学習は、FewSum の Fine-tuning に代わる処理である。

4.1 FewMoESum

FewMoESum では MoE を利用して教師なしドメイン適応を行う。図 2 に FewMoESum の概略図を示す。モデルは、事前学習で得たレビュー生成モデル (Encoder, Generator) と Plug-in Network を組み合わせ、新たに Gating モジュールを追加したものである。この際、Generator 内の Source-Target Attention モジュールをソースドメインの数と一致するよう複製し、各ソースドメインに対応したエキスパートとする。

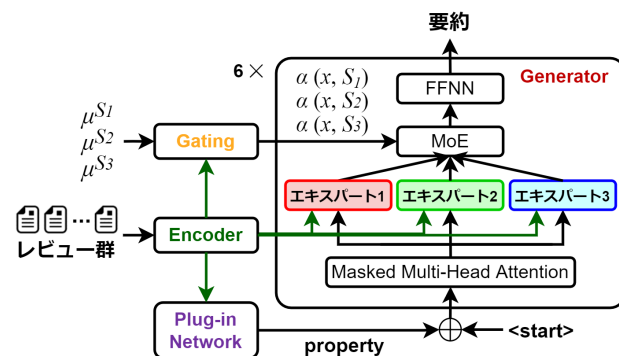


図 2 FewMoESum のモデル概略図

4.1.1 Gating モジュール

Gating モジュールは、各エキスパートの出力を加重平均する際の重みを生成する。GuoMoE と同様に、重みは入力サンプルとソースドメインとのマハ

ラノビス距離をもとに予測する。Few-Shot 学習では利用できる教師データがごく少数に限られており、大量のパラメータを調整することができない。マハラノビス距離による重み付けは大量のパラメータを必要としない(実験ではパラメータ数がモデル全体の0.1%以下)ため、Few-Shot 学習に適する。

式(1)はマハラノビス距離を示し、入力サンプルとソースドメインとの意味的な距離を予測するために用いる。次に式(2)により、求めた距離の符号を反転させ、入力サンプルとソースドメインの類似度を求める。以上の処理を各ソースドメインに対して行い、最後に、式(3)で示される Softmax 関数により各ソースドメインに対する類似度を確率で求め、各エキスパートに対する重みとする。これにより、入力サンプルとソースドメインが類似するほど、そのドメインに対応するエキスパートの出力が重視される。

$$d(x, S) = \left((E(x) - \mu^S)^T M^S (E(x) - \mu^S) \right)^{\frac{1}{2}} \quad (1)$$

$$e(x, S) = -d(x, S) \quad (2)$$

$$\alpha(x, S_i) = \frac{\exp(e(x, S_i))}{\sum_{j=1}^K \exp(e(x, S_j))} \quad (3)$$

- x : 入力サンプル
- S : ソースドメイン
- $E(x)$: 入力サンプル x の特徴量
- μ^S : ソースドメイン S の平均特徴量
- M^S : ソースドメイン S 固有のパラメータ行列
学習により S の逆共分散行列に近似する
- $\{S_i\}_{i=1}^K$: ソースドメインの集合

ただし、 $E(x)$ は入力レビュー集合に含まれる各単語の特徴量を Encoder により抽出し、平均することにより求める。

GuoMoE では、各ソースドメインに対してそれぞれ 32 個のサンプルを含むミニバッチを利用し、学習および推論を行う。この際、ミニバッチに含まれるデータを Encoder に入力することで各データの特徴量を獲得し、それらの平均を求め、各ソースドメインに対する μ^S としている。しかし、本研究では利用するモデルや入力サンプルが比較的大きく、多くの GPU メモリを消費する。そのため、ミニバッチに含まれるサンプル数が少なくなり、従来の手法を利用すると求めた μ^S の信頼性が低くなるという問題がある。したがって、FewMoESum の学習を行う直前にソースドメインの教師なしデー

タから μ^S を求め、全学習ステップおよび推論に利用する。通常、学習時に Encoder が更新されることにより、特徴空間が変化してしまうため、学習前に計算した μ^S を利用することはできない。しかし、FewMoESum の学習では、FewSum の Fine-tuning と同様に Encoder のパラメータを固定して学習するため、この手法が利用できる。

4.2 FewMoESum の学習

FewMoESum の学習では、モデルに対して要約生成への適応と、ターゲットドメインへのドメイン適応を同時に行う。この際、教師ありデータを用いて学習を行う。しかし、本手法では教師なしドメイン適応を行うため、訓練、検証にはソースドメインのデータのみを用いる。図 3 に示すように、Encoder の入力レビュー集合であり、Generator の入力はレビュー集合に対応する正解要約と、Plug-in Network により生成された property、Gating モジュールから生成された各エキスパートに対する重みである。

FewSum では、教師データの少なさに起因する過学習を抑えるため、調整するパラメータを一部に限定する手法が提案された。したがって本手法においても、調整するパラメータを Plug-in Network と各エキスパート、Gating モジュールに限定する。

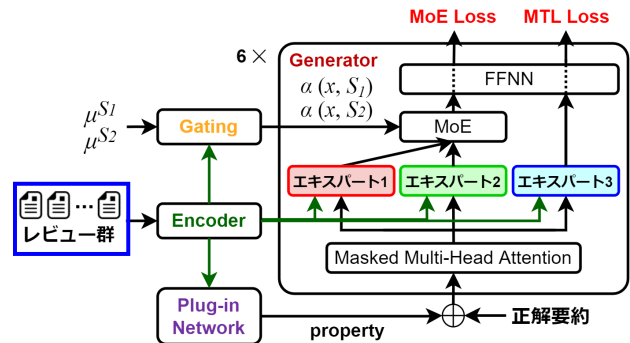


図 3 FewMoESum のモデル概略図 (学習時)

4.2.1 損失関数

学習時は GuoMoE において提案されたメタ学習手法を用いる。この手法では、エキスパートを対応するソースドメインに特化させるための $Loss_{mtl}$ と、各エキスパートに対する重み付けの学習を行うための $Loss_{moe}$ から構成される損失関数に従う。以下に損失関数を定義する。

$$Loss = \lambda \cdot Loss_{moe} + (1 - \lambda) \cdot Loss_{mtl}$$

図 3 は、エキスパート 3 に対応するソースドメインのサンプルが入力された場合の処理である。Loss_{ml} は、全エキスパートのうち、入力サンプルのドメインに対応するエキスパートのみの出力を利用して計算する。この時、Gating モジュールや他のエキスパートは使用しない。Loss_{moe} は、入力サンプルのドメインに対応するエキスパート以外の出力を利用して計算する。この時、各エキスパートの出力は Gating モジュールにより生成された重みを用いて加重平均される。また、 λ はハイパーパラメータであり、Loss_{ml} と Loss_{moe} の重みを制御する。

4.2.2 テスト

モデルのテスト時には、図 2 のように全てのエキスパートの出力を加重平均を求め、1 つの要約を生成する。入力サンプルと類似するソースドメインに対応したエキスパートの出力を重視するため、未知であるターゲットドメインのデータに対しても頑強な要約生成が可能となる。

5 実験

5.1 実験データ、評価方法

実験に使用したデータセットを表 1 に示す。データセットは FewSum の実験で利用されたものと同じであり、Amazon の製品に対するレビューと要約で構成される。なお、教師なしデータは 1 データあたり 9 個のレビュー、教師ありデータは 1 データあたり 8 個のレビューと 3 個の要約を含む。これらのデータは Clothing, Shoes and Jewelry; Electronics; Health and Personal Care; Home and Kitchen の 4 分野に分類されている。

実験時には、4 つの分野のすべての組み合わせについて、1 分野をターゲットドメイン、他 3 分野をソースドメインとしモデルの評価を行った。教師なしドメイン適応の効力を確認するため、訓練、検証時に使用する教師ありデータはソースドメインのものに限定する。テスト時は、ターゲットドメインの教師ありデータのみを用いる。また、事前学習時に用いる教師なしデータはターゲットドメインに関わらず、すべての分野のデータを利用する。

評価指標には、要約タスクにおいて一般的な自動評価である ROUGE F1 スコア [14] を使用する。ROUGE F1 スコアはテキスト間の単語の一致度を数値化する指標である。

表 1 実験に使用したデータセット

	教師なし		教師あり		
	訓練	検証	訓練	検証	テスト
Cloth	129,570	0	7	3	5
Electronics	155,549	0	7	3	5
Health	49,358	22,888	7	3	5
Home	97,721	0	7	3	5

5.2 結果

ROUGE F1 スコアを表 2 に示す。各ターゲットドメインの実験で、より優れたスコアを太字で示す。提案手法は FewSum と比較し"Health"ドメインでは精度の向上が見られたものの、ほとんどのスコアにおいて大幅な向上はなく、ドメイン適応が十分に機能していない。この原因として、教師ありデータのみを用いてドメイン適応を行ったことが考えられる。利用できる教師ありデータは少数であり、実際のソースドメインと学習に利用する教師ありデータでは分布が異なる。そのため、Gating Module がソースドメインの分布を正確に捉えられず、エキスパートに対し正しく重み付けができていないことが考えられる。同様に、エキスパートが学習に利用する教師ありデータの分布にのみ特化しており、汎化性能が低下したことが原因として考えられる。

表 2 実験結果 (ROUGE F1 スコア)

		ターゲットドメイン			
		Cloth	Electronics	Health	Home
FewSum	ROUGE-1	0.3517	0.3548	0.2937	0.3697
	ROUGE-2	0.0797	0.0813	0.0645	0.0884
	ROUGE-L	0.2176	0.2248	0.1997	0.2434
提案手法	ROUGE-1	0.3575	0.3501	0.3080	0.3596
	ROUGE-2	0.0803	0.0825	0.0689	0.0834
	ROUGE-L	0.2196	0.2290	0.2060	0.2327

6 まとめ

本研究では、Few-Shot 学習による意見要約モデルに対し、MoE を利用した教師なしドメイン適応を行う手法を提案した。評価実験では、提案手法は FewSum と比較し一部精度が向上したものの、大幅な向上には至らなかった。今後は、ドメイン分類の学習により Gating Module を初期化する、各ソースドメインで独立した追加の事前学習によりエキスパートを訓練するなど、教師なしデータを活用したドメイン適応手法について検討する予定である。

謝辞

本研究の一部は、JKA 補助事業の助成を受けたものである。

参考文献

- [1] Arther Bražinskas, et al. 2020. Few-Shot Learning for Opinion Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119-4135.
- [2] Nadav Oved, Ran Levy. 2021. PASS: Perturb-and-Select Summarizer for Product Reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 351-365.
- [3] Hayate Iso, et al. 2021. Comparative Opinion Summarization via Collaborative Decoding. arXiv:2110.07520.
- [4] Qiaozhu Mei, et al. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web in Association for Computing Machinery (ACM)*, pages 171-180.
- [5] Ivan Titov, Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web in Association for Computing Machinery (ACM)*, pages 111-120.
- [6] Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3675-3686.
- [7] Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1223-1232.
- [8] Arthur Bražinskas, et al. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5151-5169.
- [9] Ke Wang, Xiaojun Wan. 2021. TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 729-742.
- [10] Dustin Wright, Isabelle Augenstein. 2020. Transformer Based Multi-Source Domain Adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963-7974.
- [11] Jiang Guo, et al. 2018. Multi-Source Domain Adaptation with Mixture of Experts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4694-4703.
- [12] Young-Bum Kim, et al. 2017. Domain Attention With an Ensemble of Experts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 643-653.
- [13] Ashish Vaswani, et al. 2017. Attention is all you need. In *Advances in neural information processing systems (NIPS)*, pages 5998-6008.
- [14] Chin-Yew Lin. 2004. Rouge: A package for automatic eval-

uation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out Post Conference Workshop of the Association for Computational Linguistics (ACL)*, pages 74-81.