文書要約における転移学習のための文書選択手法の提案

白井　匡人† 　若林　 啓††

† 島根大学 学術研究院理工学系　〒 690-8504 島根県松江市西川津町 1060
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †shirai@cis.shimane-u.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　本研究では，対象領域の文書要約モデルの精度向上に寄与する学習文書を他の大規模な文書集合から選択する手法を提案する．転移学習はモデルの学習に対象領域以外のデータを使用する枠組みである．対象領域のデータ数が少ない場合，他の領域の文書集合から得られるデータを基に不足した情報を補うことで対象領域のモデルの精度が向上することが知られている．しかしながら，転移学習では対象領域と関連性の低いデータを学習に使用することでモデルの精度が悪化する負の転移が発生する可能性がある．提案手法は情報源となる文書集合内の各文書と対象領域の学習文書の類似度を基に学習に使用する文書を選択する．また，関連性の低い文書を対象領域のモデルの学習に用いることで負の転移が発生することを示す．

キーワード　文書要約，転移学習

1. 前　書　き

文書要約は抽出型要約と生成型要約の 2 つに分けられる．抽出型要約は本文の要約文として適切な文章を既存の文章集合から選択する．生成型要約は学習データの本文から要約文を生成する関係性を学習することでテストデータの本文を基に要約文を生成する．生成型要約は新たな文章を生成するため柔軟な要約が行えるが，適切な要約文を生成するには大量の学習データが必要となる．一般的に文書要約では要約したい対象の文書と同一の領域の文書を学習データとして使用する必要があり，任意の文書を対象に自動要約することは容易ではない．

解析対象のデータに対して十分な学習データが得られない場合，不足する学習データを補完するための仕組みとして転移学習がある．転移学習は，対象領域の解析に情報源領域から得られた情報を利用する．対象領域に少量の学習データが存在する場合，情報源領域と対象領域の学習データを基に領域の対応付けを行う．このような転移学習の設定は，帰納転移学習と呼ばれる．転移学習を用いることで対象領域から十分な学習データが得られない場合でも，情報源領域に存在する学習データをモデルの学習に使用することで有効に解析が行える可能性がある．文書要約においても関連する領域の学習データを使用することで対象領域の要約精度が上昇することが考えられる．しかしながら，転移学習では各領域の関連性が低い場合，誤った情報が転移されることで精度が悪化する負の転移が発生することが知られている [1]．このため，対象領域のデータに対して適切な学習データを選択する必要がある．

本研究では，文書要約タスクにおいて転移学習を用いた学習を行うために，3 つの問題を論じる．第 1 の問題は，文書要約タスクにおいて負の転移が発生するのか，第 2 の問題は，転移学習を行うための情報源領域の文書を如何に選択するか，第 3 の問題は異なるデータセットから転移学習を行うことで文書要約の精度を上昇させることができるかである．本研究は，対象領域の文書要約モデルの精度向上に寄与する学習文書を大規模な文書集合である情報源領域から選択する手法を提案する．提案手法は情報源となる文書集合内の各文書と対象領域の学習文書から求まる確率分布の類似度を基に学習に使用する文書を選択する．また，関連性の低い文書を対象領域のモデルの学習に用いることで要約の精度が悪化する負の転移が発生することを示す．

第 2 章では関連研究について述べ，第 3 章では提案手法について述べる．第 4 章では実験により有効性を示す．第 5 章で結論とする．

2. 関 連 研 究

帰納転移学習は，情報源領域と対象領域の共に学習データが存在する転移学習の設定である．対象領域の学習データは少量であるため，情報源領域の学習データを対象領域の特徴に合わせて適応させる必要がある．文書は話題によって使われる単語が異なるため，異なる単語分布を持つ領域であっても情報の転移が行える転移学習手法が提案されている [2] [7] [8]．帰納転移学習の先行研究として，Rai らは情報源と対称領域から初期のモデルを構築する手法を提案している [4]．この問題設定では，情報源の分布が対象領域と同一でなければならず，領域間の関連性が低い場合を想定していない．Zhu らは，シグモイド関数に基づきモデルの重みを更新する手法を提案している [5]．この手法は，情報源のデータを対象領域の学習データに直接用いている．これらの手法では，各領域の関連性が低い場合に誤った情報が伝搬され，精度が悪化する可能性がある．Chattopadhyay らは，単一の凸最適化問題を解くことによって転移学習と領域適合のための能動学習を同時に行う枠組みを提案している [6]．この手法は，初期の学習にテストデータのみしか利用できない状況に対応することができる．しかし，能動学習を行う際に情報源から学習データを 1 つずつ選択するため計算量が膨大になるという欠点がある．

近年，文書分類や固有表現抽出に転移学習が用いられ，その有効性が示されている [14] [15]．これらの研究では転移学習を用いることで精度が上昇することを示しているが，それぞれのタスクにおいて負の転移が発生することを示唆している．Tan らは情報源領域と対象領域を繋ぐ中間領域を経由することで負の転移を回避するための方法を提案している [9]．文章要約のための転移学習手法として Keneshloo らは，対象領域以外のデータセットを学習データとして使用するための手法を提案している [11]．この手法は，情報源領域の学習データを対象領域のモデルの学習に使用することで要約精度が上昇することを示している．しかしながら，領域間の関連性に着目しておらず，用いるデータセットによって負の転移が発生する可能性がある．

3. 情報源領域の文書選択手法

本研究では，生成型の文書要約モデルとして Pointer-Generator モデルを使用する [10]．提案手法では，対象領域の学習データと共に情報源領域の文書から学習に使用するデータを選択する．文書の選択には KL 情報を用い，対象領域の学習データから求まる単語の確率分布 P と類似する情報源領域の文書を選択する．

3.1 問題設定

各領域は，情報源領域 S と対象領域のテストデータ $U = \{u_1, u_2, ..., u_j\}$ と対象領域の少量の学習データ $L = \{l_1, l_2, ..., l_i\}$ とする．文書要約を行うため，対象領域の学習文書集合 L と情報源領域 S を用いて文書要約モデルを学習する．文書要約のモデルには Pointer-Generator モデルを使用する．使用する学習データを選択するために各領域の単語の確率分布を基に，KL 情報量を用いて情報源領域 S に存在する各文書 $s_1, s_2, ..., s_k$ と対象領域の学習データ $l_1, l_2, ..., l_n$ の対応付けを行う．

3.2 Pointer-Generator を用いた文書要約

Pointer-Generator モデルは，注目する箇所を表すアテンション分布と本文からコピーする箇所を決定するコピー機能を持つ要約モデルである．また，Coverage Mechanism により，表現の繰り返しを防止することができる．Pointe-Generator モデルは語彙外の単語が含まれる文書の要約に強いという特性を持つため，少量の対象領域の学習データと異なる領域のデータセットを用いる転移学習の設定に適している．

N 個の文書集合を $D = d_1, ..., d_N$ とすると，各文書は，単語数 v の次元を持つ $d_i = t_1, ..., t_v$ で表される．各エンコーダは，入力として単語 t の埋め込みを受け取り出力状態 h_t を生成する．デコーダはエンコーダから文章の最終状態である h_{t_v} を取得することで対応した出力の生成を行う．アテンションベクトル α_j，コンテキストベクトル c_j，および出力分布 p_{vocab} は次の式で計算される

$$f_{ij} = v_1^T tanh(W_h h_i + W_s s_j + b_1) \tag{1}$$

$$\alpha_j = softmax(f_j) \tag{2}$$

$$c_j = \sum_i^{T_w} \alpha_{ij} h_i \tag{3}$$

$$p_{vocab} = softmax(v_2(v_3[s_j \oplus c_j + b_2]) + b_3) \tag{4}$$

ここで，$v_{1,2,3}$，$b_{1,2,3}$，W_h，W_s は学習対象となるモデルのパラメータである．アテンション分布を持つ seq2seq モデルでは，p_{vocab} を使用してクロスエントロピー損失を求める．p_{vocab} は語彙内の単語の分布のみを扱うため，語彙外の単語は考慮できないが，Pointer-Generator モデルは $(1 - \sigma)$ の確率で単語の選択を語彙または元の文書から切り替えて選択することでこの問題を軽減している．

$$\sigma_j = (W_c c_j + W_s s_j + W_x x_j + b_4) \tag{5}$$

$$p_j^* = \sigma_j p_{vocab} + (1 - \sigma_j \sum_{i=1}^{T_e} \alpha_{ij}) \tag{6}$$

ここで，W_c，W_x，b_4 は学習対象となるモデルのパラメータであり，単語 x_j が語彙外の単語の場合，$p_{vocab} = 0$ となり，モデルはアテンション分布に基づいて元の文書から適切な単語を選択する．最終的なクロスエントロピー（CE）損失は次の式から求まる．

$$Ł_{CE} = -\sum_{t=1}^{T} logp_\theta^*(y_t|e(y_{t-1}), s_t, c_{t-1}, \mathbf{X}) \tag{7}$$

θ は学習のパラメータであり，$e(.)$ は特定の単語埋め込みを表す．

3.3 KL 情報量による類似文書の選択

情報源領域の学習データから対象領域の文書要約モデルの学習に使用するデータを選択するために KL 情報を用いる．まず対象領域の全ての学習データ $l_1, l_2, ..., l_n$ から単語の確率分布 P を求める．対象領域の学習データ集合 L の各単語の出現確率は以下の式より求める．

$$P(w|L) = \frac{n_{w,L} + \alpha}{\sum_w^V n_{w,L} + V\alpha} \tag{8}$$

ここで，$n_{w,L}$ は文書集合 L での単語 w の出現回数，V は語彙数，α はスムージングのパラメータである．対象領域の学習データから求まる単語の確率分布 P と類似する情報源領域の文書を選択する．情報源領域の学習データの各文書 $s_1, s_2, ..., s_k$ に対して単語の確率分布 P_s を求め，対象領域の確率分布 P_L と比較する．確率分布の比較には KL 情報量を用いる．KL 情報量は確率分布の差の尺度であり確率分布 P_s と確率分布 P_L が与えられたとき，KL 情報量は以下の式で求まる．

$$KL(P_s//P_L) = \sum_i P_{s,i} log \frac{P_{s,i}}{P_{L,i}} \tag{9}$$

KL 情報量の値は確率分布の差が大きいほど大きな値を取り，P_s と P_L が等しいとき KL 情報量は 0 となる．提案手法では，対象領域の文書と類似した文書を学習に使用するため，KL 情報量の値が低い上位 N 個の文書を学習に使用する．

	要約の単語数の平均	本文の単語数の平均	要約率の平均
対象領域の学習データ	41.9	653	0.0847
対象領域のテストデータ	41.9	653	0.0850
情報源領域全体	27.1	673	0.0756

表 1　各データに含まれる単語数の平均

4. 実　　験

実験では CNN データセット [13] を対象領域，CORNELL NEWSROOM データセットを情報源領域として用いる．文書選択手法の有効性を確認するため，類似度が上位 1 万，下位 1 万，ランダムに 1 万文書，情報源の全文書を追加の学習データとして使用した場合と対象領域の学習データのみを学習に使用した場合の ROUGE スコアを比較する．また，要約文の単語数に着目して情報源領域の文書を選択することで各データセットの要約文の長さの違いによる影響を調査する．

4.1 実 験 準 備

実験に用いる CNN データセットは CNN のニュース記事の本文とその要約が付随したデータセットである．CORNELL NEWSROOM データセットは 38 の出版社の著者と編集者によって書かれた 130 万の記事と要約が含まれる大規模なデータセットである．実験に用いる文書数は，対象領域の学習データとして 18516 文書，情報源領域の文書として 995040 文書を用いる．テスト文書は，対象領域の 18515 文書とする．また，提案手法は文書選択により 1 万文書を追加の学習データとして用いる．文書の前処理として，文字列は全て小文字に変換する．類似度の計算では実験データ中で出現回数が 5 以下の単語を除外して計算する．Pointer-Generator の学習の繰り返し回数は 5000 回とする．

データセットの詳細を表 1 に示す．要約率は要約文の単語数を本文の単語数で割ったものである．対象領域の学習データとテストデータはほとんど同じ傾向を示している．情報源と対象領域のデータを比較すると，本文に含まれる単語数はほとんど変わらないが，情報源のほうが要約文が短い傾向にある．要約文の付けられ方が要約の精度に影響する可能性があるため，本文に対する要約文の長さが対象領域と似た傾向になるように情報源の文書を選択した場合の精度を測定する．ここでは，対象領域の学習データの 99%が含まれる要約率が 0 から 0.24 までの範囲を 0.02 刻みでまとめ，各区間に含まれる文書の割合が同じになるように情報源の類似度上位から該当する文書を選択した場合の精度を測定する．また，要約文の単語数の度数分布が対象領域の学習データと同じになるように，対象領域の学習データの 99%が含まれる単語数 17 から 62 までの範囲を 5 刻みでまとめ，各区間に含まれる文書の割合が同じになるように類似度上位から該当する文書を選択した場合の精度を測定する．

4.2 評 価 方 法

実験の評価には ROUGE スコアを用いる [12]．ROUGE スコアは文書要約や機械翻訳の評価に使用される尺度であり，要約の正解と生成された要約を比較することで要約の精度を表す．ROUGE スコアのうち，ROUGE-N は N-gram での単語の一致を評価し，ROUGE-L は最長一致での単語の一致を評価する．ROUGE-N，ROUGE-L スコアは以下の式で求める．

$$R\text{-}N_R = \frac{Count_{match}(C(summary_{words}, reference_{words}))}{C(reference_{words})} \quad (10)$$

$$R\text{-}N_P = \frac{Count_{match}(C(summary_{words}, reference_{words}))}{C(summary_{words})} \quad (11)$$

$$R\text{-}N_F = \frac{1}{N}\sum_{i}^{N} \frac{2 * R\text{-}N_{R_i} * R\text{-}N_{P_i}}{R\text{-}N_{R_i} + R\text{-}N_{P_i}} \quad (12)$$

$$R\text{-}L_R = \frac{LSC(C(summary_{words}, reference_{words}))}{C(reference_{words})} \quad (13)$$

$$R\text{-}L_P = \frac{LSC(C(summary_{words}, reference_{words}))}{C(summary_{words})} \quad (14)$$

$$R\text{-}L_F = \frac{1}{N}\sum_{i}^{N} \frac{2 * R\text{-}L_{R_i} * R\text{-}L_{P_i}}{R\text{-}L_{R_i} + ROUGE\text{-}L_{P_i}} \quad (15)$$

Recall は真の要約をどれだけ再現できたかを表し，Precision は生成した要約がどれだけ真の要約に含まれるかを表す．f 値は Recall と Precision の調和平均である．本研究では，N-gram として 1-gram と 2-gram を用いる．

4.3 実 験 結 果

表 2，3，4 に実験結果を示す．対象領域+上位は対象領域の学習データに加えて情報源領域で類似度が上位となる 1 万文書を学習データに使用した場合である．対象領域+下位は対象領域の学習データに加えて情報源領域で類似度が下位となる 1 万文書を学習データに使用した場合である．対象領域+ランダムは対象領域の学習データに加えて情報源領域の文書をランダムに選択した 1 万文書を学習データに使用した場合である．対象領域+全文書は対象領域の学習データに加えて情報源領域の全ての文書を学習データに使用した場合である．対象領域は対象領域の学習データのみを学習に使用した場合である．上位は情報源領域で類似度が上位となる 1 万文書のみを学習データに使用した場合である．下位は情報源領域で類似度が下位となる 1 万文書のみを学習データに使用した場合である．対象領域+要約率は対象領域の学習データに加えて情報源領域で類似度が上位の文書から対象領域の学習データと要約率の分布が同じになるように 1 万文書を選択して学習データに使用した場合である．対象領域+要約単語数は対象領域の学習データに加えて情報源領域で類似度が上位の文書から対象領域の学習データと要約文の単語数の分布が同じになるように 1 万文書を選択して学習データに使用した場合である．

	$ROUGE\text{-}1_{recall}$	$ROUGE\text{-}1_{precision}$	$ROUGE\text{-}1_{fvalue}$
対象領域+上位	0.377	0.301	0.325
対象領域+下位	0.213	0.267	0.233
対象領域+ランダム	0.329	0.316	0.315
対象領域+全文書	0.306	0.301	0.297
対象領域	0.275	0.324	0.292
上位	0.345	0.309	0.317
下位	0.057	0.058	0.054
対象領域+要約率	0.272	0.319	0.288
対象領域+要約単語数	0.277	0.315	0.290

表 2　ROUGE-1

	$ROUGE\text{-}2_{recall}$	$ROUGE\text{-}2_{precision}$	$ROUGE\text{-}2_{fvalue}$
対象領域+上位	0.137	0.106	0.116
対象領域+下位	0.043	0.054	0.047
対象領域+ランダム	0.113	0.107	0.107
対象領域+全文書	0.104	0.101	0.100
対象領域	0.074	0.087	0.079
上位	0.120	0.106	0.109
下位	0.003	0.003	0.003
対象領域+要約率	0.065	0.076	0.069
対象領域+要約単語数	0.069	0.078	0.072

表 3　ROUGE-2

	$ROUGE\text{-}L_{recall}$	$ROUGE\text{-}L_{precision}$	$ROUGE\text{-}L_{fvalue}$
対象領域+上位	0.335	0.266	0.288
対象領域+下位	0.180	0.227	0.197
対象領域+ランダム	0.290	0.277	0.277
対象領域+全文書	0.269	0.264	0.261
対象領域	0.252	0.297	0.268
上位	0.304	0.272	0.279
下位	0.055	0.056	0.052
対象領域+要約率	0.247	0.289	0.261
対象領域+要約単語数	0.250	0.285	0.261

表 4　ROUGE-L

表 2 より各 f 値は, 対象領域+上位で 0.325，対象領域+下位で 0.233，対象領域+ランダムで 0.315，対象領域+全文書で 0.297，対象領域で 0.292，上位で 0.317，下位で 0.054 となる．表 3 より各 f 値は, 対象領域+上位で 0.116，対象領域+下位で 0.047，対象領域+ランダムで 0.079，対象領域+全文書で 0.100，対象領域で 0.107，上位で 0.279，下位で 0.003 となる．表 4 より各 f 値は, 対象領域+上位で 0.288，対象領域+下位で 0.197，対象領域+ランダムで 0.277，対象領域+全文書で 0.261，対象領域で 0.268，上位で 0.279，下位で 0.052 となる．ROUGE-1,ROUGE-2,ROUGE-L の f 値は，対象領域に加えて類似度の高い情報源領域の文書を使用する提案手法が最も高い値を示している．類似度が上位となる情報源領域の文書のみを学習に使用した場合の f 値が次いで高くなっており，類似度が下位となる文書のみを学習に使用した場合，要約の精度が著しく低くなる．

4.4 考　　察

表 2 より，情報源領域の文書を全て学習に使用した場合，対象領域のみを学習に使用した場合と比較して f 値が 0.5%しか向上しておらず，学習データの増加が精度の向上に寄与していない．類似度が上位となる情報源領域の学習データを使用することで対象領域の学習データのみを使用する場合と比較して 3.3%向上している．情報源領域の文書をランダムに選択した場合と比較しても 1.0%向上していることから，KL 情報量に基づき情報源領域の文書を選択する提案手法は有効であると考えられる．また，類似度が下位となる情報源領域の文書を学習に使用した場合，対象領域の学習データのみを使用する場合と比較して 5.9%精度が悪化している．このため，文書要約タスクにおいても類似度の低い情報源領域から知識の転移を行うことで負の転移が発生していると考えられる．表 3，4 より，ROUGE-2, ROUGE-L においても ROUGE-1 と同様の傾向を示しており，情報源領域の中から対象領域に対して適切な文書を選択することで要約の精度が向上している．

各データセットの要約の長さの違いによる影響を調査するために，本文と要約文の割合と要約文の単語数を考慮して情報源領域の文書の選択を行った場合の精度を測定したが，対象領域のみを学習に使用した場合より精度悪化している．要約率が同

様の分布になるように情報源を選択すると 1 万文書の中で最も類似度が低い文書は 145055 位の文書となる．このため，情報源として類似度が高くない文書が選択され，精度向上に寄与すると考えられる類似度が上位の文書の割合が減少している．Newsroom データセットは CNN データセットよりも短い要約文が付与される傾向にあるが，要約の長さを考慮せずに情報源として用いる文書を選択する方が有効であると考えられる．

5. 結　　論

本研究では，文書要約タスクにおいて転移学習を行うための文書選択手法について論じた．提案手法を用いて情報源領域の選択を行った場合，提案手法は，ROUGE-1 スコアが 3.3%改善する．また，対象領域に対して類似度の低い文書を選択した場合，負の転移が発生することを示した．これにより，提案手法の有効性を示した．

6. 謝　　辞

本研究の一部は，JSPS 科研費 (課題番号 19K20333) の助成によって行われた．

文　　献

[1] Rosenstein, M.T., Marx, Z., Kaelbling, L.P. and Dietterich, T.G.: To Transfer or Not To Transfer, In NIPS'05 Workshop on Transfer Learning, volume 898, 2005

[2] Shi, X., Fan, W., and Ren, J.: Actively transfer domain knowledge, In Proceeding of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD) 2008, pp. 342-357, 2008

[3] Raj, S., Ghosh, J. and Crawford, M. M.: An active learning approach to knowledge transfer for hyperspectral data analysis, In Proceeding of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS) 2006, pp. 541-544, 2006

[4] Rai, P., Saha, A., Daume, H. and Venkatasubramanian, S.: Domain adaptation meets active learning, In Proceeding of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing, pp. 27-32, 2010

[5] Zhu, Z., Zhu, X., Ye, Y., Guo, Y.F. and Xue, X.: Transfer active learning, In Proceeding of the 20th International Conference on Information and Knowledge Management (CIKM) 2011, pp. 2169-2172, 2011

[6] Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S. and Ye, J.: Joint transfer and batch-mode active learning, In Proceeding of the 30th International Conference on Machine Learning (ICML), pp. 253-261, 2013

[7] Shirai, M., Liu, J. and Miura, T.: Transfer Learning using Latent Domain for Document Stream Classification, In Proceedings of the Second IEEE International Conference on Multimedia Big Data (BigMM), pp. 82-88, 2016

[8] Long, M., Wang, J., Ding, G., Cheng, W., Zhang, X. and Wang, W.: Dual Transfer Learning, In Proceeding of the SIAM International Conference on Data Mining (SDM), pp. 540-551, 2012

[9] Tan, B., Song, Y., Zhong, E. and Yang, Q.: Transitive Transfer Learning, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), 2015

[10] See, A., Liu, P. J. and Manning, C. D.: Get To The Point: Summarization with Pointer-Generator Networks, In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1073-1083, 2017

[11] Keneshloo, Y., Ramakrishnan, N. and Reddy, C. K.: Deep Transfer Reinforcement Learning for Text Summarization, In Proceedings of the 2019 SIAM International Conference on Data Mining, pp. 675-683, 2019

[12] Lin, C.: ROUGE: a package for automatic evaluation of summaries, In Proceedings of ACL Workshop on Text Summarization Branches Out, 2004

[13] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P.: Teaching machines to read and comprehend. In Advances in neural information processing systems, pp. 1693-1701, 2015

[14] Semwal, T., Yenigalla, P., Mathur, G., and Nair, S. B.: A practitioners' guide to transfer learning for text classification using convolutional neural networks. In Proceedings of the 2018 SIAM International Conference on Data Mining, pp. 513-521. 2018

[15] Lin, B. Y., and Lu, W.: Neural adaptation layers for cross-domain named entity recognition, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2012-2022, 2018