転移学習を用いた参考文献書誌情報抽出の検討

Examination of bibliography extraction from reference strings using transfer learning

木下 諒
Ryo Kinoshita

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要 電子図書館の運用には，書誌情報データベースの整備が必須である．特に学術論文の参考文献欄には，著者名やタイトルといった有用な書誌情報が集約されている．しかし，CRF により書誌情報を高精度に抽出するには一定量の学習データが必要であり，その作成コストが問題になる．そこで本研究では，他雑誌で学習した書誌情報抽出機の推定結果を別の雑誌で利用する転移学習を用いて，学習データを削減する方法を検討する．

1 はじめに

膨大な文書が格納されている電子図書館を快適に利用するためには，検索やソート，文書間リンクなどの機能が必要である．しかし，これらの機能を実現するために，著者名や論文題目名などといった書誌情報を人手でデータベースに登録するコストは大きく，書誌情報を自動抽出する技術が求められている．荒内ら[1]は，自然言語処理などの様々な分野で利用されている識別モデルの一つである Conditional Random Field (CRF) を利用して，論文中の参考文献文字列から書誌情報を自動で抽出する方法を提案した．しかし，高精度に書誌情報を抽出するには学術雑誌ごとに一定量の参考文献文字列を学習データが必要であった．そのため川上ら[2]は能動サンプリング，擬似学習データの追加，転移学習を利用して，学習データを削減する方法を提案した．本研究では川上らが[2]で行った実験のうち，転移学習に焦点をあてる．具体的には，他雑誌について学習した書誌情報抽出器の推定結果を利用して，対象雑誌の書誌情報抽出器の書誌情報抽出精度をどの程度向上させることができるかを実験により確認する．

2 CRF を用いた書誌情報抽出

2.1 CRF

本研究では，標準的なチェーンモデルの CRF[3]を用いて，まず参考文献文字列をトークン列に変換する．次に各トークンに書誌要素ラベルを付与することで，書誌情報を自動抽出する．また，CRF の素性テンプレート[*1]に，47 種類の Unigram 素性と 1 種類の Bigram 素性の合計 48 種類の素性を用いる． Unigram 素性には，トークンの文字数，数字や記号などの割合などの参考文献文字列の言語的な特徴がある．また，Bigram 素性は，ラベルの連接に関する情報である．

2.2 書誌要素ラベル付与

本研究では，参考文献文字列を変換したトークン列に著者名やタイトルなどの書誌情報ラベル付与する．本稿では，CRF による書誌要素ラベル付与の精度を評価するために，人手でラベル付与されたトークン列を使用する．書誌要素ラベルは，Author や Title など 18 種類のラベルが定義されているが，評価の際には川上ら[2]の研究に倣って，表 1 に示すように，似ている要素のラベルはまとめて合計 9 クラスの分類とみなす．

2.3 転移学習

雑誌ごとに参考文献文字列の書式は異なるため，高精度に書誌情報を抽出するためには，学習データは対象雑誌のものを用いるのが通常である．ただ，参考文献文字列に表れる特徴には雑誌の種類によらない共通点が存在するため，一概に他雑誌の学習データが利用できないとはいえない．形式が類似しているデータや，言語が一致しているデータであれば，他雑誌の学習データでも役に立つ可能性がある．そこで，他雑誌で学習した書誌情報抽出機の推定結果を別の雑誌で利用する転移学習を用いて，学習データを削減する方法を検討する．

具体的には，人手で書誌要素ラベル付けした学習データを用いて，各論文誌において書誌情報抽出器を作成する．別の雑誌の書誌情報抽出器が推定した書誌要素ラベルを，対象雑誌の書誌情報抽出器の素性に加える．

表 1：書誌要素ラベルと評価時の分類クラス

書誌要素ラベル	評価時クラス
<Author>, <Editor>, <Translator>, <Author Other>	AUTHOR
<Title>, <Book Title>	TITLE
<Journal>, <Conference>	JOURNAL
<Volume>, <Number>, <Page>	VOLUME
<Publisher>	PUBLISHER
<Day>	DAY
<Month>	MONTH
<Year>	YEAR
<Location>, <URL>, <Other>	OTHER

[*1] http://taku910.github.io/crfpp/

3　評価実験

3.1　実験概要

2.3 節で説明した転移学習を行って，抽出対象雑誌の書誌情報抽出器の書誌情報抽出精度をどの程度向上させることができるかを実験により評価する．

実験には以下 2 つの論文誌の参考文献文字列コーパスを使用する．

・電子情報通信学会英文論文誌（IEICE-E）4,497 件

・IEEE Computer Society 英文論文誌（IEEE-CS）4,126 件

また，書誌情報抽出精度は 5 分割交差検定を用いて算出する．本稿では，全トークンが正しく書誌情報ラベル付けされた参考文献文字列数を全参考文献文字列数で割った値を書誌情報抽出精度とする．

3.2　転移学習の結果

転移学習による書誌情報抽出結果を図 1 に示す．書誌情報の抽出対象雑誌が IEICE-E の場合，転移学習により 0.9709 から 0.9713 へ 0.04 ポイント，IEEE-CS の場合，0.9231 から 0.9249 へ 0.18 ポイント，書誌情報抽出精度はいずれの場合もわずかながら向上した．

3.3　データを減らした実験

図 1 では，元の書誌情報抽出精度が低い雑誌の方が，転移学習の効果が大きかった．そこで，元の書誌情報抽出精度が低いほうが転移学習の効果が大きいかどうかを調べるために，IEICE-E と IEEE-CS のデータ数を 500 件にして同様の実験を行った．その書誌情報抽出結果を図 2 に示す．対象雑誌が IEICE-E の場合，転移学習により 0.948 から 0.952 へ 0.04 ポイント，IEEE-CS の場合，0.830 から 0.834 へ 0.04 ポイント，書誌情報抽出精度は向上した． しかし図 1 と比べて，転移学習の効果が高いことは確認できなかった．

4　おわりに

本研究では，他雑誌で学習して作成した書誌情報抽出器の予測ラベルを転移素性として加えて，対象雑誌から書誌情報を抽出する実験を行った．その結果，書誌情報抽出精度はわずかながら向上したが，元の書誌情報抽出精度が低い雑誌が転移学習の効果が高いことは確認できなかった．

今後の課題として，実験に使用する論文誌を増やすことや，対象雑誌の学習データのみを減らした場合の転移素性の効果を確かめる追加実験などを行いたい．また，本研究では転移素性を対象雑誌の抽出器に追加するモデルを用いたが，他の方法の転移学習を用いての参考文献書誌情報抽出の有効なモデルに関しても検討したい．さらに，転移元と転移先の論文の言語の違いによる精度の違い，形式の違いによる精度の違いなどを明らかにしたい．

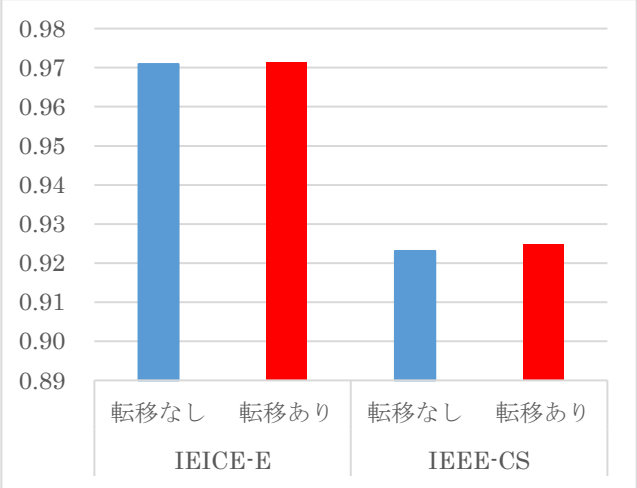

図 1：転移学習による書誌情報抽出精度

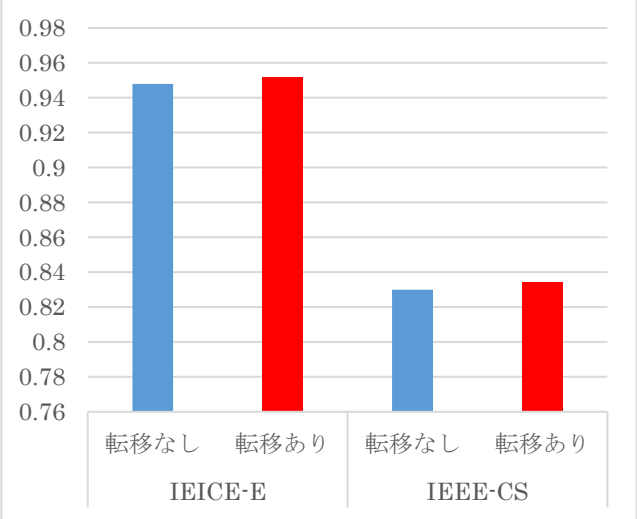

図 2：転移学習による書誌情報抽出精度（500 件）

5　参考文献

[1] 荒内大貴，太田学，高須淳宏，安達淳：CRF による和英文の参考文献文字列からの自動書誌要素抽出，情報処理学会研究報告，Vol. 2012-DBS-156，No. 1，pp. 1–8 (2012).

[2] 川上尚慶，太田学，高須淳宏，安達淳：少量学習データによる参考文献書誌情報抽出精度の向上，情報処理学会論文誌データベース，Vol. 8，No. 2，pp. 1–12 (2015).

[3] Lafferty, J., McCallum, A. and Pereira, F.: *Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data*, In Proc. of 18th International Conference on Machine Learning, pp. 282–289 (2001).