Bi-directional LSTM-CNN-CRF による参考文献書誌情報抽出の検討

Examination of Bibliography Extarction from Reference Strings by Bi-directional LSTM-CNN-CRF

浪越 大貴

Daiki Namikoshi

岡山大学 太田研究室

Ohta Laboratory, Okayama University

概要 膨大な文書が格納されている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などの有用な書誌情報が集約されているため，参考文献文字列から書誌情報を自動抽出する研究が行われている．本稿では，系列ラベリングタスクで高精度を達成している Bi-directional LSTM-CNN-CRF（BiLSTM-CNN-CRF）モデルを用いて，参考文献書誌情報抽出を行い，その抽出精度を実験により確認する．

1 はじめに

多数の学術論文を蓄積する電子図書館のサービスを利用する際，検索や文書間リンク等の機能は必須である．しかし，そのための書誌情報を人手でデータベースに入力するコストは膨大なため，その作業を可能な限り自動で行う文書解析技術が求められている．これまでに，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，参考文献文字列から書誌情報を自動抽出する研究が行われている [1]．また我々は，近年自然言語処理などのタスクで高精度を達成しているニューラルネットワーク（NN）を利用して，参考文献文字列から書誌情報を自動抽出することを試みた [2]．しかし，その抽出精度は川上らの CRF [1] に比べ，劣る結果となった．そこで本稿では，系列ラベリングのタスクで高精度を達成している Ma ら [3] の BiLSTM-CNN-CRF モデルを用いて，参考文献書誌情報抽出を行い，抽出精度を実験により評価する．

2 参考文献書誌情報抽出

2.1 問題定義

本研究では，参考文献文字列をトークン列に変換し，そのトークン列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．本稿では書誌要素ラベル付与の精度を確認するため，[1] と同様にトークン列への変換は人手で行い，トークン列への書誌要素ラベル付与は BiLSTM-CNN-CRF モデルで行う．付与する書誌要素ラベルは，Author や Title など 13 種類が定義されている（表 2）．ここで，トークンとは，BibTeX で参考文献を出力する際のエントリを構成する各文字列のことである．例えば以下の参考文献文字列は実験で使用する Cora データセット [4] のものである．

Harandi, M. T., Ning, J. Q., (1990), Knowledge Based Program Analysis, IEEE Software.

これは以下のようにトークンに分割され各トークンに書誌要素が付けられる．本来書誌要素を分けるデリミタも各書誌要素のトークンに含まれる．

<author>Harandi, M. T., Ning, J. Q., </author>
<date>(1990), </date>
<title>Knowledge Based Program Analysis,</title>
<journal>IEEE Software. </journal>

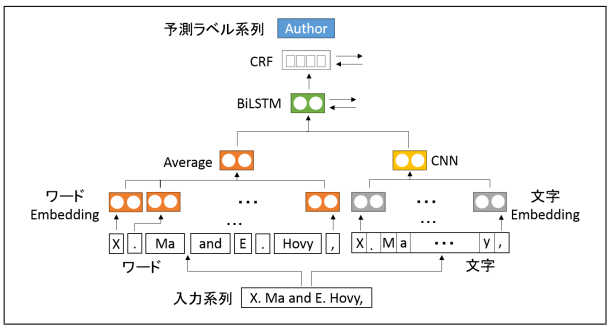

図 1: BiLSTM-CNN-CRF モデル

2.2 BiLSTM-CNN-CRF モデル

本研究で用いる BiLSTM-CNN-CRF モデルの構造を図 1 に示す．図 1 は，[3] の参考文献文字列の一部を例としたものである．BiLSTM-CNN は，トークンを構成するワードの分散表現から求めたトークンベクトルとトークンの各文字のベクトルを畳み込みニューラルネットワーク（Convolutional neural network: CNN）で畳み込んだ文字 CNN ベクトルを結合して BiLSTM 層に入力し，書誌要素を予測するモデルである．さらに BiLSTM-CNN-CRF は，Linear Chain CRF で予測するラベル系列を最適化する．

ここで，ワードの分散表現は word2vec により獲得する．word2vec のモデル作成には，実験で用いる参考文献文字列コーパスを学習データとして用いる．なお，カンマなどのデリミタも文脈として考慮するため，前処理としてあらかじめ定義されているデリミタ [1] の前後に半角スペースが存在しない場合，半角スペースを挿入してから学習する．モデル作成後，そのトークンに含まれるワードのベクトルの相加平均をワードから求めたトークンベクトルとする．ここで，ワードは，トークンを半角スペースで区切った文字列である．例えば，図 1 に示した [3] の著者である “X. Ma and E.

表 1: 各モデルの書誌情報抽出精度

	書誌情報抽出精度
BiLSTM-CNN-CRF	0.874
BiLSTM-CRF	0.788
CRF [1]	0.850
2BiLSTMs-CNN [2]	0.798

表 2: 各書誌要素の正解率

書誌要素	要素数	BiLSTM-CNN-CRF	CRF [1]	2BiLSTMs-CNN [2]
Author	489	0.9939	1.0000	1.0000
Booktitle	230	0.9348	0.9130	0.8870
Date	495	0.9899	0.9677	0.9576
Editor	43	0.8837	0.7209	0.6977
Institution	58	0.8966	0.9310	0.9310
Journal	165	0.9333	0.9394	0.8970
Location	137	0.9562	0.9562	0.9197
Note	30	0.9333	0.8667	0.7667
Pages	288	0.9826	0.9792	0.9792
Publisher	101	0.9802	0.9901	0.9109
Tech	61	0.8852	0.9016	0.8852
Title	493	0.9878	0.9939	0.9919
Volume	181	0.9834	0.9890	0.9503
Average	2,771	0.9726	0.9675	0.9516

Hovy,”というトークンは，“X”，“.”，“Ma”，“and”，“E”，“.”，“Hovy”，“,”という 8 つのワードに分割される．よって，このトークンのワードから求めるトークンベクトルはこの 8 つのベクトルの平均となる．また，文字 CNN ベクトルは入力トークンの文字列の各文字のベクトルを CNN に入力として与えた固定次元のベクトルである．ここで，CNN に入力する文字ベクトルは，keras [*1]に実装されている Embedding レイヤーを用いて獲得する．

3 評価実験

3.1 実験概要

BiLSTM-CNN-CRF モデルによる抽出精度の評価のため，研究論文の 19,890 件の英文の参考文献文字列からなる Cora データセット [4] を用いる．学習データを 19,390 件，テストデータを 500 件とする．

評価指標として参考文献文字列を構成する全トークンに正しく書誌要素ラベルを付与できた参考文献文字列数を，全参考文献文字列数で割った書誌情報抽出精度を算出する．また書誌要素毎の正解率も求める．

3.2 実験結果

本稿では，BiLSTM-CNN-CRF モデルを川上らの CRF [1] と [2] の NN モデルを比較する．[2] の NN モデルは 2 つの BiLSTM 層と 1 つの CNN 層をもつため，2BiLSTMs-CNN と表記する．

[*1] https://keras.io

ここでトークンベクトルの獲得に利用する word2vec では，minCount を 0，次元数は 100 とした．また，文字ベクトルの獲得に利用する Embedding 層では，次元数を 40 とした．BiLSTM-CNN-CRF の CNN では，畳み込みフィルターの数を 30，1 フィルターがカバーする文字数は 5 とした．BiLSTM 層では，出力系列の次元を 80 に設定した．最適化関数は rmsprop，バッチサイズは 10，学習回数は 100 回に固定した．また，文字 CNN ベクトルの有効性を検証するため，図 1 の右側の文字 CNN ベクトルを除いたモデル（BiLSTM-CRF）とも比較する．

各モデルの書誌情報抽出精度を表 1 に示す．表 1 より，BiLSTM-CNN-CRF モデルは最も抽出精度が高く，参考文献書誌情報抽出のタスクにおいても有効であった．また，BiLSTM-CRF モデルよりも抽出精度が 8.6 ポイント高いことから，文字の畳み込みが抽出精度の向上に寄与している．また，各モデルの書誌要素毎の正解率を表 2 に示す．ここで，表 2 の“Average”は全参考文献文字列の全トークンに対し，正しくラベルを付与できたトークンの割合（マイクロ平均）を表す．表 2 より，“Booktitle”や“Editor”のような，それぞれ“Title”，“Author”と区別がつきにくい書誌要素に対しても，2 つの比較対象と比べ，正解率が高い．しかし，“Title”，“Author”の正解率は 2 つの比較対象と比べ高くない．また，“Institution”や“Tech”のような，参考文献文字列に表れにくい書誌要素に対しても，正解率が高くない．そこで，このような書誌要素にも対応できるように，モデルの拡張やデータの前処理方法を検討したい．

4 まとめ

本研究では BiLSTM-CNN-CRF モデルを用いて，参考文献書誌情報抽出を行い，先行研究の CRF と NN によるモデルと書誌情報抽出精度を比較した．その結果，BiLSTM-CNN-CRF モデルはこれらよりも抽出精度が高く，参考文献書誌情報抽出のタスクにおいても有効であることを確認した．今後の課題として，BiLSTM-CNN-CRF モデルの拡張方法の検討，他の参考文献文字列コーパスにおける実験が挙げられる．

参考文献

[1] 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. 18-29, 2015.

[2] 浪越大貴，太田学，高須淳宏，安達淳，“分散表現と素性を利用した参考文献書誌情報抽出,” DEIM Forum 2018, I5-1, 2018.

[3] X. Ma and E. Hovy, “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF,” in Proc. of ACL 2016, pp. 1064-1074, 2016.

[4] A. McCallum, K. Nigam, J. Rennie and K. Seymore, “Automating the Construction of Internet Portals with Machine Learning,” Information Retrieval, vol. 3, no. 2, pp. 127-163, 2000.