ブートストラップ法による科学ニュース記事からの雑誌名抽出

菊地 真人　　吉田 光男　　梅村 恭司
豊橋技術科学大学 情報・知能工学系
{m143313@edu, yoshida@cs}.tut.ac.jp, umemura@tut.jp

1 はじめに

日本語の科学ニュース記事では，研究成果がわかりやすく述べられるが，出典となる文献情報は明記されない傾向にある．このことは，読者が研究の詳細を知ることへの障壁となっている．一方，研究内容が掲載された雑誌名は記事中に明記されることが多く，雑誌名を自動抽出することで対象の文献情報を探索する手がかりが得られる．そこで我々は，日本語の科学ニュース記事からの雑誌名抽出に取り組み，得られた雑誌名をリスト化する．このリストは，記事から雑誌名を識別するための辞書として利用でき，コンパクトで扱いが容易と考える．

文書から雑誌名などの固有表現を自動抽出するタスクとして固有表現抽出がある．このタスクでは，固有表現がタグ付されたデータセットを用いた教師有り学習手法がしばしば用いられる．固有表現を学習する外部知識として，固有表現辞書の利用も有効である．辞書利用の利点は高い抽出精度であり，固有表現を十分被覆する大規模な辞書を用意することで多くの固有表現を高精度に抽出できるが，そのような辞書を人手で用意することはコストが高い．雑誌名を抽出対象とする場合も以下の理由から，辞書を人手で用意することは困難である．雑誌名は多種多様である．また，外国語の雑誌名を日本語の記事で掲載する場合は外来語として扱わず，発音に文字を当てはめて翻訳することが多いため，記事を書いた著者によって雑誌名の表記ゆれが起こる．そのため我々は，ブートストラップ法 [1, 2, 3] に基づいて科学ニュース記事からの雑誌名を抽出することにした．ブートストラップ法は，人手で付与したシードと呼ばれる少数の固有表現を教師データとして，固有表現の抽出と辞書の拡充を交互に繰り返す．この方法によって，少数の固有表現をもとに多くの固有表現を抽出できる．

雑誌名は果たす役割が定まっており，特定の文脈に出現しやすい傾向がある．それゆえ，雑誌名抽出では分布仮説 [4] の適用が有効と考えた．分布仮説は，「同じ文脈で使われる語彙は，類似する意味を持つ傾向にある」というものである．そこで本稿では，雑誌名抽出に対する分布仮説の有効性を検証する．この検証のために，雑誌名の出現パターンとして左右の文脈のみを手がかりとした抽出モデルを構築し，ブートストラップ法を用いて雑誌名抽出を試みる．雑誌名抽出の手がかりとして，あえて左右の文脈のみを使用することで，雑誌名抽出に対する分布仮説の影響を正確に分析できる．

本稿の貢献は，雑誌名が特定の文脈に出現しやすいという仮定を立て，雑誌名抽出に対してこの仮説を裏付けたことである．雑誌名抽出の結果として，分布仮説は雑誌名の特定に重要な役割を果たすことが示唆された．さらに，抽出結果の失敗例についても分析し，ブートストラップ法によって継続的に精度よく雑誌名を抽出するためには，左右の文脈に加えて他の情報も必要となることを示した．

2 関連研究

ブートストラップ法を用いて固有表現を獲得する代表的手法として，次の先行研究がある．Riloff らは，固有のパターンとドメイン特有の語彙の同時抽出，そして曖昧性の高い語彙を辞書から振るい落とす 2 層のブートストラップ法を提案した [1]．Collins らは，固有表現のスペリング，出現文脈を利用して固有表現を分類した [2]．Yangarber らは，固有表現の周囲から固定長の文脈を参照し，パターンとして利用した [3]．ただし，固有表現抽出のパターンマッチに利用するパターンは左右の文脈のうち，どちらか一方である．本研究でもブートストラップ法を採用し，周囲の文脈からパターンを学習する．ただし，提案する雑誌名抽出は，雑誌名の両側から学習した文脈をパターンとして利用するところが特徴である．

日本語では，パターンとして両側の文脈が用いられる [5, 6]．雑誌名は構成要素数が多く，付属語や活用

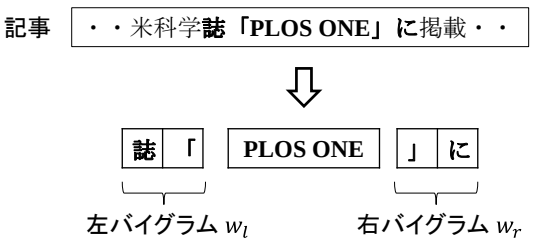

図 1: 左右バイグラムの抽出例

語を含む場合がある．手法 [5, 6] では，形態素解析を使用するため，しばしば付属語や活用語を含む固有表現を抽出できない．それゆえ，我々は形態素解析の代わりに文字ベースの枠組みを使用する．

文書／文レベルですべての単語を素性とした分布類似度を用いたアプローチが提案されている [7]．しかし，すべての単語を素性とすると，その素性空間は多次元かつ疎となるという問題がある．我々は，雑誌名抽出において分布仮説が十分有効に働くと考え，雑誌名の周囲にある文脈のみを利用することとした．

3 文字ベースの枠組み

提案手法は，記事本文を入力として，本文に含まれる雑誌名をブートストラップ方式で抽出する．

手順 1. 訓練データの作成：まず，記事本文を文字バイグラム単位に分割する．データセット全体でバイグラムの種類ごとに出現頻度を数え上げ，頻度情報を保持する．続いて，データセットから辞書にある雑誌名を最長一致検索ですべて発見し，雑誌名の左右にあるバイグラムを抽出する（図 1）．これらのバイグラムについて，種類ごとに出現頻度を数え上げ，左右別々に頻度情報を保持する．最後に，数え上げたバイグラムの頻度を部分毎（データセット全体，雑誌名の左右）に Simple Good-Turing 法 [8] により補正し，それを確率の推定に用いた．我々は，頻度補正によってゼロ頻度問題に対処することができ，雑誌名抽出の性能も向上すると考えた．

手順 2. 雑誌名の抽出：抽出対象とする雑誌名の文字列長を範囲指定し，データセットから指定した長さの雑誌名候補および左右の文字バイグラムを抽出する．抽出の際は，雑誌名候補の開始位置を記事の先頭に固定し，候補の長さを指定した範囲の中で一文字ずつ大きくしながらがら候補と左右バイグラムを抽出していく．候補の長さが範囲の最長となった，あるいは候補が記事末尾に達した後は，開始位置を一文字だけ右にシフトして再度抽出を始める．

抽出した雑誌名候補についてスコアを計算する．雑誌名候補 w のスコアは下式に示す尤度比の相乗平均とした．

$$S(w) = \left\{ \frac{P(w_l \mid O_l)}{P(w_l \mid O_{all})} \times \frac{P(w_r \mid O_r)}{P(w_r \mid O_{all})} \right\}^{\frac{1}{2}}$$

候補の左右バイグラムをそれぞれ w_l，w_r とする．O_{all}，O_l，O_r はバイグラムがそれぞれデータセット全体，雑誌名の左，右で出現することを表す．各尤度比は，バイグラムが特定部分（雑誌名の左右）に出現する確率とデータセット全体に出現する確率の比で表される．それぞれの確率は，手順 1 で保持した観測頻度を用いて最尤推定し，推定した確率の比を取ることで尤度比を計算する．

スコアの降順から上位 N 件を抽出後，雑誌名か否かを人手で判定し，雑誌名であったものを辞書に追加する．

手順 3. 繰り返し：十分な数の雑誌名が得られるまで手順 1 と 2 を繰り返す．

4 評価実験

ウェブから収集した科学ニュース記事から雑誌名を抽出し，提案手法の性能を定量的に評価する．シードとして表 1 に示す 10 個の雑誌名を使用した．抽出対象となる雑誌名の長さは，雑誌名の判定が困難な 1 文字を除き，2 文字から 50 文字までとした．雑誌名候補のうち，スコアの降順上位 2,000 件（$N = 2,000$）を抽出し，これらの候補を抽出雑誌名と定義する．性能の評価尺度は適合率，部分再現率および F 値である．いずれの実験とも雑誌名の抽出を 2 回までを評価する．すなわち，まずシードを用いて雑誌名を抽出し，次にその雑誌名を辞書に加えて，再び雑誌名を抽出する．上位 2,000 件にはシード以外の雑誌名候補を含め，再度抽出を行う場合はシードおよび過去に抽出した候補以外を上位 2,000 件に含める．なお，辞書に雑誌名を加える際は，シードをもとに抽出した上位 2,000 件のうち，人手で雑誌名と判定したもののみを辞書に加える．これによって，雑誌名ではない文字列の影響を取り除き，分布仮説の有効性を分析する．

4.1 データセット・雑誌名について

学術雑誌名を含む可能性の高い科学ニュース記事をデータセットとして使用する．具体的は，複数の日本語ニュースサイトから過去およそ 10 年分のニュース

表 1: シードとして使用する雑誌名

Journal name
Scientific Reports
サイエンティフィック・リポーツ
サイエンティフィック・リポーツ（Scientific Reports）
サイエンティフィックリポーツ
サイエンティフィックリポーツ（Scientific Reports）
PLOS ONE
プロス・ワン
プロス・ワン（PLOS ONE）
プロスワン
プロスワン（PLOS ONE）

表 2: 雑誌名の表記パターン

英名 例： Neuron，Nature
和名 例： ニューロン，ネイチャー
英名・和名の併記 例： ニューロン（Neuron），ネイチャー（Nature）
補足情報付き 例： ニューロン電子版，ネイチャー（電子版）

記事を収集し，「学誌 OR 論文誌 OR 学術誌」という検索条件で絞り込んだ合計 30,076 記事を使用した．雑誌名抽出の際は，記事本文のみを参照する．雑誌名の抽出元が日本語記事であるため，和文雑誌および英文雑誌の名称が多く抽出される．雑誌名の表記は，大別して表 2 に示すパターンがある．

4.2 正解データと評価尺度

性能を測る尺度として次に示す適合率，部分再現率および F 値を用いる．

$$\text{適合率} = \frac{\text{人手で正解と判定した抽出雑誌名の数}}{\text{これまでに抽出した抽出雑誌名の数}}$$

$$\text{部分再現率} = \frac{\text{正解データが含む抽出雑誌名の数}}{\text{正解データが含む雑誌名の数}}$$

$$\text{F 値} = \frac{2 \cdot \text{部分再現率} \cdot \text{適合率}}{\text{部分再現率} + \text{適合率}}$$

再現率を計算する際には，正解データとして人間が雑誌名と認識できる全文字列の集合が必要である．しかし，このような集合を作成することはできない．そこで，Web of Science，ScienceDirect に収録されている雑誌名リスト，および国会図書館が索引を作成している雑誌名一覧を取得し，これらの雑誌名をもとに部分再現率を計算する．取得した雑誌名のうち，データセットに含まれる雑誌名を最長一致検索で調べ，含まれない雑誌名は削除した．日本語の雑誌名には，一般的な用語と区別しにくい雑誌名が含まれる．そのため，

表 3: 性能評価の結果（数値はすべて累積）

Metric	Number of iterations	
	1	2
適合率	0.856	0.636
部分再現率	0.517	0.587
F 値	0.645	0.611
雑誌名数	1,712	2,544

IPA 辞書 [1] に含まれる名詞全般と雑誌名の一致検索を行い，一致した雑誌名は削除した．それでも一般的な用語と区別しにくい雑誌名が多く含まれていたため，残っているすべての雑誌名から 9 文字以下の雑誌名を一律に取り除き，正解データとした．この正解データは，1,037 件の雑誌名を含む．この正解データは，記事に含まれるすべての雑誌名を包含するわけではないが，機械的に作成できる範囲において定量的な性能評価の近似値を得ることができる．

4.3 実験結果

文字ベースの枠組みについて，性能評価の結果を表 3 に示す．反復の 1 回目では，適合率が 0.8 以上と高く，部分再現率も 0.5 を超えていた．このことから，少数の雑誌名をシードとして与えた場合でも，左右の文脈をパターンとすれば多くの雑誌名を抽出できることが分かった．しかし，反復の 2 回目では新たに得られた雑誌名が半減し，適合率が 0.2 ほど低下したにもかかわらず部分再現率はほとんど向上しなかった．この結果から，雑誌名を抽出する手がかりとして左右バイグラムを使うことは有効であるが，ブートストラップ法で高い F 値を維持したまま，雑誌名を抽出するためには工夫が必要ということが示唆された．

5 考察

反復の 2 回目において学習した左右バイグラムのうち，高頻度の上位 3 件について頻度および左右バイグラム全体に占める出現割合を調べた結果を表 4 と表 5 に示す．上位 3 件のバイグラムだけを見ても，それらの出現割合は左右バイグラムそれぞれで全体の 50%，25%以上を占める．このことは，雑誌名が限られた文脈に出現する傾向にあることを数値的に示している．そのため，雑誌名抽出の際は，雑誌名周囲の文脈情報

[1] IPA 辞書として，mecab-ipadic-2.7.0-20070801 を用いた．

表 4: 高頻度の左バイグラム上位 3 件と出現割合

左バイグラム	補正頻度	出現割合
学誌	14,237.146	0.334
誌「	9,735.146	0.229
ー（	1,112.148	0.026
合計	**25,084.440**	**0.589**

表 5: 高頻度の右バイグラム上位 3 件と出現割合

右バイグラム	補正頻度	出現割合
」に	4,450.283	0.105
に発	4,206.283	0.099
電子	2,239.284	0.053
合計	**10,895.850**	**0.257**

表 6: 誤った抽出の例

例	左バイグラム	文字列	右バイグラム
1	咳）	ワクチンを	接種
2	学誌	ネイチャーで 13 日	に発
3	誌「	ジャーナル・オブ	・パ

をうまく捉えることで，多くの雑誌名を抽出できると考える．

実験での誤抽出について傾向を述べる．反復の 2 回目における誤抽出の例を表に示す．誤り例 1 は，曖昧性のある雑誌名「ワクチン」から文脈を学習してしたために，雑誌名とは無関係の文字列を抽出してしまったものである．このように，抽出したい意味カテゴリと無関係の文字列を獲得してしまうことは意味ドリフトと呼ばれる．誤り例 2 では，左右の文脈は雑誌名らしいが，抽出した文字列に「で 13 日」という不要な情報が付いている．今回，誤りの大半は例 2 と同様のものであり，これを回避するためには抽出した文字列の内部も考慮して雑誌名らしさを計算する必要がある．最後に誤り例 3 は，雑誌名の一部のみが獲得された例である．雑誌名は，部分文字列として別の雑誌名を含む入れ子構造であることが多い．部分文字列となっている雑誌名から文脈を学習すると，分布仮説が成り立たないケースを生み，例 3 のように雑誌名の一部を抽出してしまうことがある．

6　まとめと今後の課題

本稿では，少数のシードから出発し，日本語の科学ニュース記事から文脈情報だけを使用して雑誌名を抽出するタスクに取り組んだ．実験では，雑誌名の出現文脈から雑誌名らしさを尤度比で推定する枠組みを構築し，科学ニュース記事から雑誌名の抽出を試みた．結果として，雑誌名は分布仮説が強く働くことが分かり，少数の雑誌名をシードとした場合であっても，文脈情報を活用することで多くの雑誌名を抽出できることが示唆された．しかしながら，文脈のみでは雑誌名の一部分や不要な文字列が付いた雑誌名が抽出されたり，意味ドリフトを起こしたりするケースも見られた．今後は，上記の問題解決を図るとともに，分布仮説に基づいた既存手法でも雑誌名の抽出性能を検証したい．

参考文献

[1] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI*, pp. 474–479, 1999.

[2] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *EMNLP*, pp. 100–110, 1999.

[3] R. Yangarber, W. Lin, and R. Grishman. Unsupervised learning of generalized names. In *COLING*, pp. 1–7, 2002.

[4] Z. S. Harris. Distributional structure. *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

[5] M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *NAACL–HLT*, pp. 8–15, 2003.

[6] K. Nakano and Y. Hirai. Japanese named entity extraction with bunsetsu features. 情報処理学会論文誌, Vol. 45, No. 3, pp. 934–941, 2004.

[7] P. Pantel, et al. Web-scale distributional similarity and entity set expansion. In *EMNLP*, pp. 938–947, 2009.

[8] W. A. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, Vol. 2, No. 3, pp. 217–237, 1995.