非即時的なタスク設定における固有表現抽出の改善

澤山 熱気 [1]　鈴木 潤 [2,3]　進藤 裕之 [1,3]　松本 裕治 [1,3]

[1] 奈良先端科学技術大学院大学 情報科学研究科
[2] NTT コミュニケーション科学基礎研究所
[3] 理化学研究所 革新知能総合研究センター
{sawayama.atsuki.ro2, shindo, matsu}@is.naist.jp
suzuki.jun@lab.ntt.co.jp

1 はじめに

インターネットの普及以降，膨大な数の科学技術論文を誰でも手軽に獲得できるようになった．また，最新の研究成果が日々投稿され公開されている．例えば，生命科学やバイオメディカル分野の論文検索エンジンで知られる Pubmed [1]のデータベースである MEDLINE は，公開されている情報に基づくと，おおよそ 40 万もの論文が毎年登録されている．これは 1 日あたり平均して 1000 本を超える計算になる．このような状況から，たとえその分野の専門家（或いは研究者）であっても，これらの大量の論文すべての内容に目を通し，理解することは非常に困難な状況にある．つまり，専門家に必須となる，分野の動向や最先端の研究成果を把握し続ける行為に多大な労力が必要となる．

この問題を緩和する方法として，自動知識抽出システムの利用が考えられる．実際に，これまでもいくつかの自動知識抽出システムが提案されてきた [1, 2] ．こういったシステムを利用することで，大量の論文から自分が必要とする情報を比較的容易に獲得することができるようになり，専門家の情報収集の労力が大幅に軽減されることが期待できる．

このような動機から，本稿では科学技術論文からの知識抽出に関して，更なる抽出精度の向上に向けた方法論の議論を行う．ただし，自動知識抽出システムは，固有表現抽出 (Named Entity Recognition, NER) モジュール，関係抽出 (Relation Extraction, RE) モジュール，辞書構築モジュールといった多くのモジュールで構成される．その全てを一度に議論するのは困難であるため，本稿では，最初の取り組みとして固有表現抽出モジュールの抽出精度向上に焦点をあてる．これは，固有表現抽出モジュールは自動知識抽出システムの最初の処理にあたり，その性能が後続のモジュールの性能に大きく影響を与えると考えられるためである．

[1]https://www.ncbi.nlm.nih.gov/pubmed

2 科学技術論文からの知識抽出タスクにおける固有表現抽出

固有表現抽出とは，文中にある専門用語などの事前に固有表現（Named Entity, NE）と定義した重要語句を抽出することである．近年では，リカレントニューラルネットワーク（Reccurent Neural Network, RNN）をベースとした条件付き確率場 (Conditional Random Field, CRF)[3, 4] が，CoNLL-2003 shered task datasets [2]などの標準的な固有表現抽出のベンチマークデータセットにおいて，高い抽出精度を示している．本稿でも，RNN による条件付き確率場をベースライン手法と想定して議論を行う．

以下，まず科学技術論文（以降「論文」と略記）からの知識抽出に向けた固有表現抽出の特性を考え，その特性に即して論文からの知識獲得に向けた固有表現抽出法を提案する．ここでは特に以降に述べる「頻出する規則性のある新しい用語」と「非即時的なタスク設定」の二点に着目する．

2.1 頻出する規則性のある新しい用語

論文からの知識抽出を想定した固有表現抽出では，比較的規則性のある専門用語が固有表現として多く出現する．例えば，バイオ分野や化学分野におけるタンパク質名や化学式などが挙げられる．また，こういった専門用語は日々新しく生み出される状況にある．こういった新しく生み出された用語は機械学習の文脈では未知語となる．一般的に，未知語は抽出精度の悪化に大きく影響を与えると考えられるため，これらの未知語をうまく取り扱う方法論が求められる．

近年，盛んに研究が進んでいるニューラルネットワークを用いた言語処理においては，未知語を扱う方法の一つとして文字列情報を組み込む方法がしばしば用いられている．本稿では，前述したように論文中には規則性のある専門用語が固有表現として出現しやすいという性質に着目し，その分野の用語として高頻度で出現する特有の部分文字列に分解して利用する方法も合わせて活用することとする．

[2]https://www.clips.uantwerpen.be/conll2003/ner/

2.2 非即時的なタスク設定

論文からの知識抽出を想定した固有表現抽出では，論文の収集・知識抽出から利用者がシステムを利用するまでの間に時間的な猶予がある．そのため，利用されるまでの時間的な猶予を活用し，獲得した論文全てを用いて繰り返し学習を行うことが可能である．つまり，学習の際に評価対象のデータを既に得ている前提で，評価データを学習に用いることができる．これは，従来，固有表現抽出タスクの利用シーンの想定となっている対話・QA(Question answering) システム等のリアルタイム性を求められる状況とは異なる．本稿では，このタスク設定を「非即時的なタスク設定」と呼ぶ．

抽出対象となるデータを学習時に既に得ているのであれば，学習用のデータだけではなく，評価データから得られる情報を有効に活用して固有表現抽出モデルを学習する方法論を選択することができる．例えば，機械学習の研究分野では，評価データを活用する学習法はトランスダクティブ学習（Transductive learning）と呼ばれ，多くの研究が行われている [5]．

トランスダクティブ学習では，モデルに与えられる評価データの正解は既知ではない．そのため，単純な自己学習では，固有表現抽出モデルが予測した固有表現タグの結果を，正誤に関わらず，正解であるものとして再学習に活用するため，再学習によって新しい情報がほとんど得られないことが容易に予想される．加えて，一度間違えてしまった固有表現タグをモデルが間違い続ける問題が発生する．このことから，再学習の際に，予測した固有表現タグ結果をそのままモデルに与えるのではなく，何らかの別の情報量としてモデルに与える方法論が求められる．

3 提案手法

3.1 部分単語への分割

2.1 節で述べた頻出する規則性のある新しい用語に適した方法として，固有表現抽出の前処理として，与えられた文章の各単語に対して部分単語（Subword）による分割を行う．このとき，部分単語への分割方法にはいくつかの方法論が考えられる．例えば，もっとも単純には人手作成のルールによる方法が考えられる．しかし，新しい用語の増加によるルールの網羅性の低下，少数の例外処理，ルール追加のメンテナンスコストなどが必要となり，あまり良い方法とは言えない．

近年，統計量に基づく部分単語への分割方法がニューラル機械翻訳の研究分野でよく用いられるようになってきた [6]．本稿では，統計的な方法を用いて部分単語への分割を行う．

一点注意として，固有表現抽出では，機械翻訳のデータとは違い固有表現タグと単語間の対応関係を部分単語へ分割する際に保持する必要がある．つまり，単語を部分単語へ分割，及び，逆変換に相当する元の分割単位に復元する際に，タグの対応関係も同様に変換可能する必要がある．しかし，これは単純な規則で容易に変換可能であるため，部分単語への分割を固有表現抽出へ適用することは問題とはならない．例として，タンパク質の固有表現の分割処理を図 1 に示す．

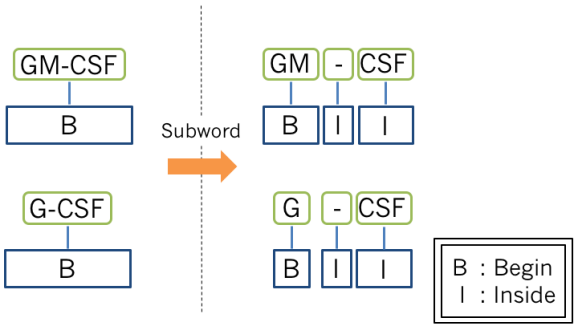

図 1: 部分単語による単語の部分文字列への分割と固有表現タグの分割例

3.2 評価データの予測分布の活用

提案手法では，2.2 節で議論した性質を効果的に利用するため，予測した固有表現タグを正解としてモデルに与える代わりに，予測分布を追加の特徴として活用する方法を考える．提案手法の基本となるアイディアは，入力データに対するタスク依存の類似度を計算し，追加の特徴として用いる方法となる．基本的な学習の枠組みは一般的な自己学習と似ているが，自己学習と提案手法との大きな違いは，モデルが予測した固有表現をどのように再学習に用いるかである．

以下学習の枠組みを説明するために，まず，提案手法の概要図を図 2 に示す．提案手法では，最初に初期ベースモデル（Initial base model, Ib-model と呼ぶ）を一般的な固有表現抽出と同じ設定で構築する（図中 1）．次に，Ib-model を用いて，与えられた評価・開発データに対してタグに対する予測分布を計算する（図中 2）． その後，予測分布を追加情報とし評価・開発データを用いて，類似度素性モデル（Similerity feature model, Sf-model と呼ぶ）を学習する（図中 3）．次に，Sf-model を用いて，学習データの類似度素性 (similerity feature) を算出する（図中 4）．最後に，類似度素性を活用して Ib-model を再学習した次期ベースモデル（Next base model, Nb-model と呼ぶ）を構築する（図中 5）．

Ib-model は，一般的な固有表現抽出モデルと同様のネットワーク構造で構築し，評価・開発データに対して固有表現タグの予測分布を計算する役割を担っている．

Sf-model は，データ間の単語それぞれの類似度を計算する役割を担っている．これによって，入力された単語が固有表現となる単語と似ているのか，固有表現となる単語と似ていないのかを類似度素性という形で算出する．次に，Sf-model における，類似度素性の算

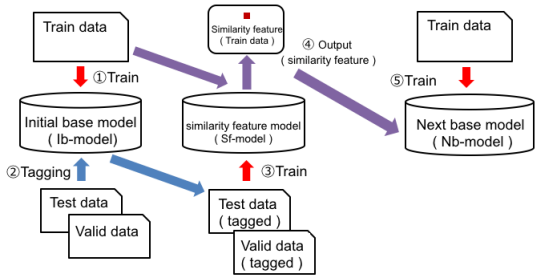

図 2: 提案手法の概略図

出方法について説明する．一般的な固有表現抽出モデルでは，各固有表現クラスの確率値を出力し，その値のうちの最も高い確率のタグを予測タグとして用いている．提案手法の類似度素性として，各固有表現タグの確率値である BIOES [7] の 5 次元の値を用いた．注意すべき点として，Sf-model では点ごとの推定結果を算出するために，CRF を用いず，RNN でのモデル構築を行う．

Nb-model では，Ib-model に Sf-model から算出した類似度素性を活用し，再学習を行う．提案手法の計算式を式 1 に示す．単語ベクトル $\boldsymbol{X}$ を Nb-model の入力として与え，RNN の線形変換後の出力値である BIOES の 5 次元の値 $\Phi(\boldsymbol{X})$ に対して，同様に Sf-model から算出した類似度素性 $\Psi(\boldsymbol{X})$ を足し合わせる．その値を $\boldsymbol{O}$ とし，CRF への入力とする (式 1)．パラメータ α は類似度素性の用いる度合い，ϵ は log スケールの補正のために導入している．

$$\boldsymbol{O} = \Phi(\boldsymbol{X}) + \alpha \log(\Psi(\boldsymbol{X}) + \epsilon) \quad (1)$$

ただし，α は [0,1.0] から設定し，$\epsilon = 1.0 \times 10^{-6}$ とする．

4 実験

本稿では，バイオインフォマティクス分野の論文からの知識抽出を想定する．ここでは論文中からタンパク質の固有表現を抽出する実験を行った．実験に用いるデータセットとして，BioNLP 2011 shared task の Entity relation supporting task データセット [8][3]を用いた．データセットは論文アブストラクトで構成されており，学習データ (800 本)，開発データ (160 本)，評価データ (250 本) から成り立つ．

4.1 部分単語による前処理

予備調査として，実験に用いるデータセットの語彙の比較を行った．表 1 に各データ間を比較した語彙の異なりを示す．この表からデータ間の語彙が大きく異なることがわかる．よって，部分単語を用いて評価データの未知語を特有の部分文字列へ分解して既知の単語として扱う必要があり，これによって，抽出精度の向上に繋がる可能性がある．

[3]http://2011.bionlp-st.org/home/entity-relations

表 1: 部分単語化によるデータの語彙の比較

	標準	Sub 1k	Sub 2k	Sub 3k
学習データ	16918	1172	2189	3193
開発データ	6009	1130	2058	2878
評価データ	9432	1163	2136	3048
学習・評価の比較				
学習・評価の和集合	21176	1177	2196	3205
学習・評価の積集合	5174	1158	2129	3036
学習データのみ出現	11744	14	60	157
評価データのみ出現	4258	5	7	12
学習・開発の比較				
学習・開発の和集合	18931	1173	2191	3201
学習・開発の積集合	3996	1129	2056	2870
学習データのみ出現	12922	43	133	323
開発データのみ出現	2013	1	2	8

表 2: 部分単語化によるデータごとの固有表現の平均単語数の変化

	標準	Sub 1k	Sub 2k	Sub 3k
学習	1.34	3.46	2.86	2.47
開発	1.32	3.40	2.88	2.49
評価	1.32	3.84	3.12	2.76

部分単語へ分割するために既存のライブラリである，Subword-NMT [6] を利用した．学習データの単語の語彙が，1000, 2000, 3000（Sub 1k，2k，3k と略記する）程度になるよう分割し，その分割方針を用いて，評価データ，開発データも同様に分割した．

データセットに対して部分単語を適用した結果を表 1，表 2 に示す．部分単語を用いたことで，各データセット間の語彙の異なりは低下し，固有表現の平均単語数は増加した．このことから，評価データに出現する未知語を低減できたことが分かる．

4.2 分割されたデータによる固有表現の抽出

固有表現抽出モデルとして NeuroNER [9] を用いた．このモデルのネットワークは Bi-directional LSTM-CRF で構築されている．実験には，NeuroNER の標準パラメータを用い，標準で用いられている学習済みの単語埋め込みベクトルのみ使用しなかった．これは，バイオインフォマティクス分野のテキストで学習されていないベクトルであること，データセットに部分単語を適用する効果を明確にするためである．Ib-model では，通常の固有表現抽出モデルと同様に NeuroNER [9] を学習し，開発・評価データにタグをつけた．その後，タグをつけた開発・評価データを用いて Sf-model を Bi-directional LSTM でモデル学習したのち，各データセットの類似度素性を算出した．Nb-model では，学習済みの Ib-model に対して類似度素性を追加して再学習を行った．このとき，Nb-model で新たに追加したパラメータ α は，[0,1.0] の値を手動で決定し

表 3: Subword-NMT 適用による効果 (F 値の変化)

	標準	Sub 1k	Sub 2k	Sub 3k
学習	99.98	99.94	99.97	99.92
開発	80.58	80.45	78.62	78.69
評価	77.42	**80.94**	79.56	78.38

表 4: 類似度素性による F 値の変化

	標準	Sub 1k	Sub 2k	Sub 3k
Ib-model	77.42	80.94	79.56	78.38
Ib-model を追加学習	77.60	81.65	79.51	78.32
Nb-model(ベスト)	78.02	**82.28**	79.80	79.06

て用いた．実験では，それぞれの単語の出力に BIOES タグを適用するため，タグの種類は 5 種類となる．実験の性能評価には，フレーズごとの F 値 [10] を用い，同一設定で Ib-model を追加学習したモデルと精度比較を行った．

4.3 結果と考察

部分単語に分割したデータセットを用いた実験結果を表 3 に示す．学習・開発データの F 値は，学習 (100 エポック) 中で最も高い値である．一方，評価データの F 値は，開発データが最も高い値となったエポックのモデルを用いて評価した際の結果である．部分単語へ分割しない設定（表中の標準）よりも部分単語を用いた設定のほうが F 値が高い結果となった．このことから，固有表現の単語数が増えることによる難しさよりも，未知語であることの難しさの方が上回るのではないかと考えられる．

次に類似度素性を用いた提案手法の結果を表 4 に示す．表は学習中で開発データの F 値が最も高かったエポック時の評価データの F 値である．表内の (ベスト) は，α の各設定中で，最も開発データが高かったエポック時の評価データの F 値を比較し，最も高かった α の設定（標準: 1.0，Sub 1k: 0.2，Sub 2k: 0.2，Sub 3k: 0.2）の F 値である．提案手法では，Ib-model よりも抽出精度の改善が見られた．このことから，評価データから得られた類似度素性がモデルの改善に役立ったことが分かった．

今回の手法の改善点として，以下の三つの点が挙げられる．一つ目は，専用の単語埋め込みベクトルを用意することである．バイオインフォマティクス専用のベクトルを用意することで精度の改善に繋がると考えられる．二つ目は，Sf-model の構築である．今回の Sf-model の構築には評価データと開発データを用いたが，より正しい類似度を計算するために，少量の評価データだけではなく，大量の論文を用いてモデル学習が必要であると考えられる．三つ目は，Nb-model の改善である．類似度素性の与え方は他にも考えられる．例えば，単語埋め込みベクトルに素性を連結する方法などが挙げられ，別の手法でも試すことで，より効果のある活用方法を検討すべきである．加えて，今回の実験ではパラメータ α を手動で決定していたため，より高い精度になるように α の最適値をモデルが自動学習できるようにする必要がある．

5 おわりに

本稿では，科学技術論文からの知識抽出を想定した固有表現抽出の精度改善のために，二つの提案を行った．一つ目は，前処理として部分単語へ分割する方法である．これはタンパク質の固有表現によく現れる部分文字列を捉えること，未知語を低減する目的で，部分単語によるデータセットの部分文字列への分割を行った．二つ目は，評価データを活用する方法である．提案手法では，評価データを用いて各データの類似度を計算し，その類似度を追加の特徴として学習する方法を用いた．実験では，これら二つの手法それぞれで，現在，固有表現抽出で高い抽出精度を示しているニューラルネットによる条件付き確率場による手法の抽出精度を向上することができた．

参考文献

[1] Kenji Yoshimura, Toru Hitaka, Sho Yoshida, et al. Automatic extraction system of technical terms in japanese science and technology sentences. *Journal of Information Processing Society of Japan*, Vol. 27, No. 1, pp. 33–40, 1986 (in Japanese).

[2] Fuchun Peng and Andrew McCallum. Information extraction from research papers using conditional random fields. *Information processing & management*, Vol. 42, No. 4, pp. 963–979, 2006.

[3] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[4] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[5] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, Vol. 1. Wiley New York, 1998.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[7] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147–155. Association for Computational Linguistics, 2009.

[8] Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. Overview of the entity relations (rel) supporting task of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 83–88. Association for Computational Linguistics, 2011.

[9] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2017.

[10] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pp. 127–132. Association for Computational Linguistics, 2000.