The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

複数サブワード系列を考慮した BiLSTM-CRF モデルを用いた文書からの化合物名抽出

関根裕人 *1 浦澤合 *1 乾孝司 *1 岩倉友哉 *2
Hiroto Sekine Go Urasawa Takashi Inui Tomoya Iwakura

*1筑波大学大学院/理研 AIP-富士通連携センター
*2富士通研究所/理研 AIP-富士通連携センター

In this paper, we propose a BiLSTM-CRF model for extracting compound names from documents in chemical domain. The proposed model can be taken multiple subword sequences as input in order to obtain sufficient features for long span or unknown tokens. Subword LSTM units with contextual information are introduced in the input layer of the model. We conducted experiments based on CHEMDNER challenge to investigate the effectiveness of the model. As a result, the extraction accuracy outperformed the normal BiLSTM-CRF model, and experimental results on unknown words showed that the proposed method works better.

1. はじめに

1.1 研究背景

人間が読める言語で書かれた文書から自動的に構造化データを抽出するタスクを「情報抽出」という。近年、Bioinfomarmatics 分野では、この情報抽出を利用した、化合物名抽出という研究分野がある。テキストから化合物名を抽出しデータを構造化することで、論文の検索性を向上させたり、データを分析し化合物間の関係について分析することができる。例えば、化合物名抽出の競争的イベントの一つに CHEMDNER[1] がある。このタスクでは PubMed という化合物関連の論文サイトの論文をアノテーションしたコーパスを作成し、コーパスとして公開している。そして、このコーパスを使用した研究活動が活発に行われている。

上記の CHEMDNER で扱われる化合物名抽出には、一般的な固有表現抽出よりも難しい点がいくつか存在する。化合物名は一つのエンティティの長さが長いものがある。例えば、“3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide”は抽出すべき一つのエンティティであるが、このエンティティは全部で 58 文字から構成されている。次に、同一のエンティティでも複数の名前がつけられていることである。フェニルアラニンがその一例である。フェニルアラニンの他に、“L-Phenylalanine”,“Phe-OH”,“Antibiotic FN-1636 ”など 30 種類以上の呼び名が存在する。また、未知語が多いということも難しい点である。化合物は日々新しく作られていて、それに伴い化合物名も増加していく。未知語に対してどの程度抽出できるかという点も重要な評価の要素となる。

1.2 目的

近年、固有表現抽出の分野では Long Short Term Memory Network(LSTM) を用いて抽出されることが多い。LSTM を使用すると単語の意味や文脈の関係を考慮した計算が可能となる。本稿では LSTM をベースとして、上記の問題を解決するためにサブワードの情報をモデルに組み込む手法を提案する。サブワードとは一つの単語をより細かい単位に分ける考え方である。化合物名には極端に長い単語があるので、サブワード化により細かく区切ることで単語のみの時よりも高い抽出精度が達成できると期待できる。

連絡先: 関根裕人，筑波大学院システム情報工学科，sekine@mibel.cs.tsukuba.ac.jp

1.3 本論文の構成

本稿ではまず、第 2 章で LSTM による固有表現抽出の関連研究をいくつかあげる。第 3 章ではベースラインとなる BiLSTM-CRF モデルについて説明する。第 4 章ではサブワードの分散表現の獲得手法について述べる。そして第 5 章で実験とその結果について述べる。最後に第 6 章でまとめを行う。

2. 関連研究

近年の化合物名抽出タスクでは固有表現抽出を系列ラベリング問題に落とし込み、ニューラルネットワークを用いて解く手法が中心である。その中でも Long Short Term Memory(LSTM) を用いたものが多い。

Jie[2] らは Neural Network による系列ラベリングの手法をまとめ、その抽出精度を測った。比較対象は LSTM や CNN などのモデルによる違いや、学習パラメータの違いや、文字系列の有無による違いなどをまとめている。この実験では固有表現抽出に対しては、BiLSTM-CRF に文字系列 LSTM を加えた結果が最も値がよかった。

Luo ら [3] は既存の BiLSTM-CRF に加えて、Attention 層を追加したモデルを提案した。Attention 層では、global vector という文全体の類似度を考慮することで、より幅の広い時系列の情報も取り込むことができるようになっている。

Akbik[4] らは BiLSTM-CRF モデルに加えて文字系列 LSTM を拡張した。一般的に文字系列の LSTM は一つの単語内に適用されるが、文全体の中で文字系列 LSTM を計算することによって、文脈を残しながら文字をベクトルに埋め込むことができると提案している。

このように単語および文字の情報を考慮した手法は多く提案されているが、本研究のようにサブワードを考慮した手法は提案されてない。

3. ベースライン手法

3.1 BiLSTM-CRF モデル

BiLSTM-CRF モデルについて説明する。モデルの全体構成を図 1 に示す。入力する単語系列、ラベル系列をそれぞれ $x = (x_1, x_2, ..., x_t, ..., x_N)$、$y = (y_1, y_2, ..., y_t, ..., y_N)$ とする。x_t は t 番目の単語の分散表現、y_t はその単語に対応するラベルを表している。

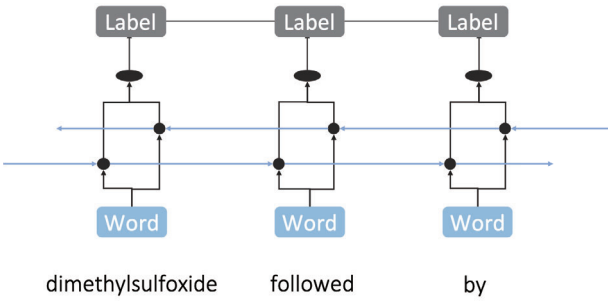

図 1: BiLSTM-CRF モデルの全体図

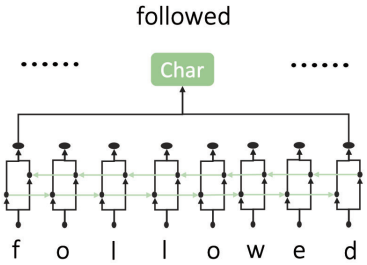

図 2: 文字 BiLSTM モデル

モデルはまず与えられた入力に対し、LSTM 層で計算を行う。LSTM とは時系列ニューラルネットワーク (RNN) の一種である。一つ前のステップの出力に加えて、時系列内で重要と考えられる情報をゲート構造を用いて保持する長期記憶と呼ばれる隠れ層を保持している。これにより一般的な RNN よりも長期的な依存関係を考慮して計算している。

$$h_t, c_t = f_{LSTM}(x_t, h_{t-1}, c_{t-1}; \theta) \tag{1}$$

BiLSTM-CRF モデルでは LSTM を前と後ろからの両方向から計算する。前向きの LSTM の出力を $\overrightarrow{h}$、後向きの LSTM の出力を $\overleftarrow{h}$ とする。これらの計算結果を各ステップごとにつなぎ合わせ、活性化関数の tanh をかける。

$$out = \tanh([\overrightarrow{h} \oplus \overleftarrow{h}]) \tag{2}$$

最後に CRF 層で、それぞれのラベル列の遷移確率を考慮し、入力系列 x に対してもっともらしい y を求める。学習時は以下の $P(y|x)$ を最大にするように、パラメータを更新する。

$$P(y|x) = softmax(Score(x, y)) \tag{3}$$

この時の $Score$ 関数は、各ラベルごとの遷移確率を $T[y_{t-1} \to y_t]$ とすると、以下の式で表すことができる。

$$Score(x, y) = \sum_{t=1}^{N} (\log(out_t) + \log(T[y_{t-1} \to y_t])) \tag{4}$$

3.2 Character Representation

単語系列に加えて、文字系列の分散表現を BiLSTM-CRF モデルに追加することで、抽出精度が上がることが多い。特に未知語に対しては、単語では情報がなくなってしまう場合でも、文字を使用することで情報を得ることができる。今回は文字系列の埋め込みを LSTM を使用して獲得する。

図 2 の部分が文字系列 LSTM の概要である。一つの単語を文字に分解し、文字単位で BiLSTM 層で計算する。前向き LSTM では一番最後の隠れ層の出力、反対に後向き LSTM では一番先頭の出力をそれぞれ連結させる。 最終的に、このベクトルと単語の分散表現に連結して、BiLSTM-CRF モデルの入力として使用する。

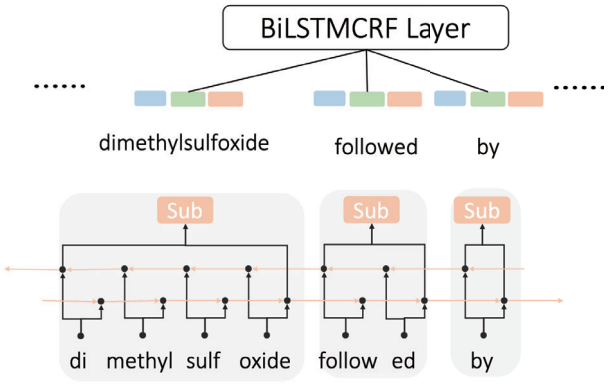

図 3: SubWord LSTM モデルの全体図

4. Subword Representation

サブワードの分散表現の獲得方法について述べる。サブワードとは、単語をさらに分割することでより意味のある情報を得ようとする考え方である。例えば“dimethylsulfoxid”という語は“di, methyl, sulf, oxid”という分割をすると“2”を意味する“di”や、“メチル基”を表す“methyl”、“酸”を表す“oxid”などの情報を得ることができる。

4.1 Subword LSTM

サブワードの分散表現を獲得するための新しいモデルを提案する。このモデルの全体図を図 3 に示す。本手法は Akbik[4] から着想を得て、サブワードを考慮するように変形しものである。

ある単語 x_t が m 個のサブワードに分割された時、$x_t = s_{x_t,1}, s_{x_t,2}, ..., s_{x_t,m_t}$ と表す。このとき、入力 $x = (x_1, x_2, ..., x_N)$ から得られるサブワード系列は $S = (s_{x_1,1}, ..., s_{x_1,m_1}, s_{x_2,1}, ..., s_{x_N,m_N})$ と表せる。

得られた S に対し、第 3 章で述べた BiLSTM 層と同様に計算する。今回、サブワード系列のベクトルを Word 系列に合わせる必要がある。前向き LSTM の計算結果では単語の最後のサブワード、後向き LSTM の計算結果では単語の先頭のサブワードに対応するベクトルを選択する。

ここで、前向き LSTM の出力を $\overrightarrow{h_t} = f_{LSTM}(s_t)$ とする。これを Word 系列と同じ長さにするためには、$(s_{x_1,m_1}, s_{x_2,m_2}, ..., s_{x_N,m_N})$ と対応する $\overrightarrow{h_t}$ を前向き LSTM の計算結果 $\overrightarrow{h_w}$ として使用する。

反対に、後向き LSTM を Word 系列に連結するためには、$(s_{x_1,1}, s_{x_2,1}, ..., s_{x_N,1})$ と対応する $\overleftarrow{h_t}$ を後向き LSTM の計算結果 $\overleftarrow{h_w}$ として使用する。

表 1: 実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.9031	0.8578	0.8799
+ SW2k(93ep)	0.9047	0.8584	0.8809
+ SW4k(82ep)	0.9032	0.8589	0.8805
+ SW16k(57ep)	0.8998	0.8577	0.8783
+ SW4k,16k(86ep)	0.9006	0.8668	0.8834
+ SW2k, 4k,16k(80ep)	0.9025	0.8566	0.8790

表 2: IV に対する実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.9095	0.8941	0.9018
+ SW2k(93ep)	0.9127	0.8921	0.9020
+ SW4k(82ep)	0.9102	0.8921	0.9011
+ SW16k(57ep)	0.9089	0.8917	0.9002
+ SW4k,16k(86ep)	0.9097	0.8976	0.9036
+ SW2k, 4k,16k(80ep)	0.9090	0.8940	0.9014

表 3: OOV に対する実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.8663	0.6871	0.7664
+ SW2k(93ep)	0.8628	0.6939	0.7692
+ SW4k(82ep)	0.8630	0.6894	0.7665
+ SW16k(57ep)	0.8619	0.6876	0.7649
+ SW4k,16k(86ep)	0.8635	0.7065	0.7771
+ SW2k, 4k,16k(80ep)	0.8507	0.6995	0.7677

最後に、前向き LSTM と後向き LSTM の計算結果をつなぎ合わせた $[\overrightarrow{h_w} \oplus \overleftarrow{h_w}]$ がサブワードの分散表現となる。このベクトルは、文字の分散表現と同様に、BiLSTM-CRF モデルへの入力に連結されて使用される。

このモデルは複数のサブワード系列を考慮する場合でも使用できる。その場合は、Subword LSTM をサブワード系列の数だけ用意し、一つの場合と同様に分散表現を得て、BiLSTM-CRF モデルへの入力として、サブワード系列の数だけベクトルを連結し使用する。

5. 評価実験

5.1 実験内容

提案モデルの有効性を調査する。ベースラインとして BiLSTM-CRF に文字 LSTM を加えたものを使用する。このベースラインに Subword LSTM を加えたときの抽出精度の有効性を調査する。

5.2 データ

BioCreative Challenge から出された Chemdner コーパス [1] を実験用のデータとする。このコーパスは PubMed 中の論文の abstract を 10,000 件集め、それらに化合物と判断したエンティティを人手でアノテーションしたものである。全部で 84,355 のエンティティが存在し、それらのユニークな数は 19,806 である。データ数は訓練、検証、テスト用それぞれ 3,500、3,500、 3,000 件ずつ提供されている。また本研究では、BIOES スキーマに従ってラベルづけを行った。

事前学習用のコーパスとして化学系の論文を扱うサイトである PubMed から Chemdner タスクに合うように選択された約 440 万件の abstract を使用した。単語系列、サブワード系列それぞれの埋め込み層の学習には GloVe[5] を使用した。

また、このコーパスを使用して SentencePiece[6] の学習も行った。学習にはユニグラムを使用し、語彙数は 2,000、4,000、16,000 の 3 つを使用した。

5.3 実験パラメータ

今回の実験では最適化に SGD を使用し、学習率は 0.005、減衰率は 0.0001 とした。単語系列、文字系列、サブワード系列の分散表現の次元はそれぞれ、50 次元、30 次元、50 次元とし、LSTM の隠れ層では 200 次元、50 次元、50 次元とした。

また、GPU は Tesla V100-DGXS を用いて学習した。バッチサイズが 10 で、100 エポック学習させた時に検証用データに対してもっとも値の良いモデルを使用した。

5.4 実験結果

実験結果を表 1 にまとめた。語彙数が 2000、4000 のサブワード系列をそれぞれひとつずつ加えた場合、ベースラインよりも F 値を上回った。これより、ベースラインにサブワード系列を追加した場合、抽出精度が向上することがある。

今回の実験で最も良い F 値であったのは、語彙数が 4,000、16,000 のサブワード系列を同時に加えた場合であった。ベースラインと比較すると、0.003 上回っており、0.8834 であった。これは、複数のサブワード系列を同時に加えることで、抽出精度が良くなる場合があることを示している。

サブワードの有効性を調べるために未知語の単語に対する結果も調査した。抽出すべきエンティティが全て未知語だった場合のエンティティを OOV(Out of vocaburaly) と呼び、反対に未知語以外の語が一つでも入っている語を IV(In vocaburaly) と呼ぶ。テストデータの全エンティティ25,308 個に対して、IV は 20,859 個で、OOV は 4,449 個であった。OOV の例としては“HANPs”, “nitriles”, “fidarestat”, “flavonolignans”, “SFN”, “CDDP”, “CYN” などがあげられる。

結果を表 2 および表 3 に示す。表から、IV に対しての結果はあまり変化がみられなかった。しかし、OOV に対してはサブワードを追加したモデルがベースラインよりも Recall が上回っていた。これは、未知語に対してサブワードの埋め込みベクトルがうまくはたらいていることを示している。また OOV のときでは、SW4k,16k はベースラインよりも 0.01 以上も F 値を上回っていた。OOV において、ベースラインでは抽出できなかったが、SW4k,16k で抽出できた例としては、“HANPs”、“nitriles”、“fidarestat”、“inaclotide”、“silatrane” などがあげられる。

6. まとめ

本稿ではテキストからの化合物名抽出において、サブワードの埋め込みベクトルをモデルに加えることで抽出精度が上がることを示した。

今回は語彙数が 2,000、4,000、16,000 に限定し実験を行ったが、それぞれで抽出の精度が異なっていた。サブワードの語彙数を決定することは、このモデルの重要な要素の一つでもあるため、今後は語彙数と抽出精度の関係性について深く調べる必要がある。

また、今回の実験では語彙数が 4,000 と 16,000 の二つのサブワード系列を入力した場合に最も良い精度となった。しかし、そのモデルに対し、語彙数が 2,000 のサブワード系列を

加えた場合の抽出精度は芳しくなかった。これは、単純にサブワード系列を加えていくのではなく、どのサブワード系列を使用するか判断する必要があることを示している。

今後は、加えるサブワード系列の語彙数と、どの語彙数を加えると良いスコアを得るかについてもう少し、研究していくことが必要である。

参考文献

[1] Krallinger et al. (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform. 2015 Jan 19;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S2. doi: 10.1186/1758-2946-7-S1-S2. eCollection 2015.

[2] Jie Yang et al. (2018) Design Challenges and Misconceptions in Neural Sequence Labeling. 2018 13 Aug. CoRR. abs/1806.04470

[3] Ling Luo et al. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018 Apr 15;34(8):1381-1388.

[4] Alan Akbik et al. (2018) Contextual String Embeddings for Sequence Labeling. 2018 Aug. Proceedings of the 27th International Conference on Computational Linguistics, p.16381649

[5] Jeffrey Pennington et al. (2014) GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP). p1532–1543. https://github.com/stanfordnlp/GloVe

[6] Taku Kudo et al. (2018) SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. 2018 Aug. CoRR. abs/1808.06226. https://github.com/google/sentencepiece