Contextual Subword Embeddingsを考慮した文書からの化合物名抽出実験

関根裕人 †* 浦澤合 † 乾孝司 † 岩倉友哉 §

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

代表連絡先：sekine@mibel.cs.tsukuba.ac.jp

1 はじめに

近年、化学物質や薬品などの化合物に関する学術論文が多く出版されている。それに伴い、多くの学術論文からテキストマイニングや情報抽出の技術を用いて、自動的に化合物名を抽出する需要が高まっている。学術論文から化合物名を自動的に抽出できるようになると、論文の化合物名によるインデックスを作成できたり、化合物間の関係を推測するのに役立てたりすることができる。

化合物名抽出は一般的な固有表現抽出と違い、難しい点が存在する。それは極端に長い単語が存在するということと、未知語が多いことである。

このような特徴より、本研究では複数のサブワード系列を用いた化合物名抽出方法を提案する。化合物名はIUPAC[8] などで決められた命名規則を用いて構造を元に名前が決められているものが多い。その命名規則によると、“methyl”や “amino”や “oxid”など、化合物名によく使用される接尾辞や接頭辞がつくことが多い。本研究では化合物名をより細かいサブワードに分割した情報を化合物名抽出に利用する。

2 関連研究

Akbik らはニューラルネットワークによる固有表現抽出の拡張として、Contextual String Embeddings[1] を提案している。これは文全体の文字系列の言語モデルの情報をを、単語系列の入力として使用するモデルである。

入力する文字系列を $X = (x_1, x_2, x_3, \cdots, x_n)$ と表す。t 番目の文字の生成確率はそれまでの文字列によって決定されるので、$P(x_t|x_{\leq t-1})$ である。$x_{\leq t-1}$ は $(x_1, x_2, ...x_{t-1})$ を表す。x_{t-1} までの系列全体の生成確率に対数をとったものを対数確率と呼び、$\sum_{t=1}^{n} logP(x_t|x_{\leq t-1})$ と表す。言語モデルではこの対数確率を最大化するように学習することで、次の文字を予測できるようになる。Contextual String Embeddings では対数確率の $x_{\leq t-1}$ の部分に、BiLSTM 層の出力を用いる。BiLSTM 層による計算は、

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

$$\overleftarrow{h_t}, \overleftarrow{c_t} = f_{LSTM}(x_t, \overleftarrow{h_{t+1}}, \overleftarrow{c_{t+1}}; \theta) \quad (2)$$

となる。ここで、$\overrightarrow{h_t}$、$\overrightarrow{c_t}$ は前向き LSTM 層の t での状態ベクトル、$\overleftarrow{h_t}$、$\overleftarrow{c_t}$ は後向き LSTM 層の t での状態ベクトルを表している。この出力は単語系列と接続して使用するため、単語系列長と合わせる必要がある。前向き LSTM の計算結果では単語の次にくる区切り文字、後ろ向き LSTM の計算結果では単語の前にある区切り文字に対応するステップをそれぞれ選択する。選択したベクトルを各ステップごとにつなぎ合わせ、活性化関数をかけたものを $h = \tanh([\overrightarrow{h} : \overleftarrow{h}])$ と表す。求めた h に対し以下の計算をする。

$$word^{Akbik} = \frac{exp(Vh + b)}{\sum_{t=0}^{n} exp(Vh + b)} \quad (3)$$

V と b はそれぞれモデルの重みとバイアスである。$word^{Akbik} = (word_1^{Akbik}, word_2^{Akbik}, ..., word_n^{Akbik})$ は文字系列 X を BiLSTM 層でエンコードしたものである。これを用いて言語モデルを作成する場合、対数確率は $\sum_{t=1}^{n} logP(x_t|word_{\leq t-1}^{Akbik})$ として表され、これを最大化するように学習する。

Akbik らは Contextual String Embeddings を学習させた後、それによって得られた分散表現である $word^{Akbik}$ と単語の分散表現の二つをつなぎ合わせたものを BiLSTM-CRF モデルの入力として使用した。

3 提案手法

本章ではサブワードを用いた BiLSTM-CRF モデルを提案する。基本的な構造は 2 章で述べた Akbik ら [1] の手法と同じである。我々の提案手法ではそれに加えて単語をさらに細かく区切ったサブワード系列

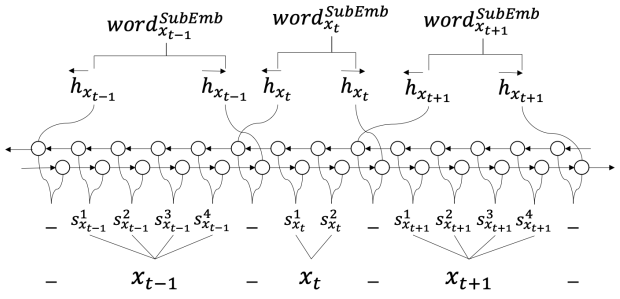

図 1: Contextual Subword Embeddings モデル

の情報を、BiLSTM を用いた言語モデルを用いてサブワードの分散表現を獲得し、単語系列の BiLSTM-CRF モデル [7] の入力としたものである。本稿ではこのサブワード系列の BiLSTM の言語モデルを Contextual Subword Embeddings と呼ぶ。

3.1 Contextual Subword Embeddings

Contextual Subword Embeddings は Contextual String Embeddings を文字系列の言語モデルからサブワード系列の言語モデルへ変更したものである。つまり、文全体のサブワード系列の言語モデルを学習し、その言語モデルから単語系列の入力となる分散表現を求める。Contextual Subword Embeddings の全体図を図 1 に示す。

ある単語 x_t が m 個のサブワードに分割された時、$x_t = s^1_{x_t}, s^2_{x_t}, ..., s^{m_t}_{x_t}$ 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 $S = (s^1_{x_1}, ..., s^{m_1}_{x_1}, s^1_{x_2}, ..., s^{m_2}_{x_2}, s^1_{x_3}, ..., s^{m_n}_{x_n})$ と表せる。得られた S に対し、2.1 節で述べた式 (1)、式 (2) と同様の計算をする。

$$\overrightarrow{h_t}, \overrightarrow{c_t} = f_{LSTM}(s_t, \overrightarrow{h_{t-1}}, \overrightarrow{c_{t-1}}; \theta) \quad (4)$$

$$\overleftarrow{h_t}, \overleftarrow{c_t} = f_{LSTM}(s_t, \overleftarrow{h_{t+1}}, \overleftarrow{c_{t+1}}; \theta) \quad (5)$$

ここで、t はサブワード系列 S のステップ数である。この出力は単語系列と接続して使用するため、単語系列長と合わせる必要がある。そこで、前向き LSTM の計算結果では単語の次の区切り文字、後向き LSTM の計算結果では単語の前の区切り文字に対応するベクトルを選択する。選択したベクトルを各ステップごとにつなぎ合わせ、活性化関数をかけたものを $h = tanh([\overrightarrow{h_t} : \overleftarrow{h_t}])$ と表す。h に対し以下の計算をする。

$$word^{SubEmb} = \frac{exp(Vh + b)}{\sum_{t=0}^{n} exp(Vh + b)} \quad (6)$$

V と b はそれぞれモデルの重みとバイアスである。$word^{SubEmb} = (word^{SubEmb}_1, word^{SubEmb}_2, ..., word^{SubEmb}_n)$ はサブワード系列 S を単語系列にあわせてエンコードしたものである。これを用いて言語モデルを作成する場合、対数確率は $\sum_{t=1}^{n} logP(x_t | word^{SubEmb}_{\leq t-1})$ と表され、これを最大化するように学習する。

Contextual Subword Embeddings を学習させた後、それによって得られた分散表現である $word^{SubEmb}$ と 2 章で述べた $word^{Akbik}$ 及び、単語の分散表現とつなぎ合わせたものを BiLSTM-CRF モデルへの入力として使用する。

3.2 複数サブワード系列を考慮したモデル

単語レベルの BiLSTM-CRF モデルに複数のサブワード系列の分散表現を入力に加えるモデルを作成することで、既存の評価セットと同様の評価をすることができ、モデルの拡張も容易に行える。そのため、複数のサブワード系列を考慮する場合でも使用できる。その場合では前節の $word^{SubEmb}$ を複数のサブワード系列によりいくつか作成し、それら全てを $word^{Akbik}$ 及び、単語の分散表現とつなぎ合わせたものを BiLSTM-CRF モデルへの入力として使用する。

複数のサブワード系列を用意することでより多くのサブワード情報を使用できる “isopropylamine” という化合物を例にとると、“iso_pro_pyl_am_ine” と区切る場合と、“iso_propyl_amine” と区切る場合とでは、後者の方がプロピル基を表す “propyl” と、アミンを表す “amine” などが獲得できるため、複数の分割方法を用いることで抽出性能がよくなると考えられる。

このように Contextual Subword Embeddings を複数用意し、様々なサブワード系列の分散表現を BiLSTM-CRF モデルの入力とすることで、より情報量のあるベクトルを入力に加えられると考えられる。

4 評価実験

4.1 データセット

BioCreative Challenge から出された CHEMD-NER コーパス [4] を実験用データとする。このコーパスは PubMed 中の論文 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。本研究では、このデータに対し BIOES 法に従ってラベルづけを行った。

また、化学系の論文を扱うサイトである PubMed から CHEMDNER タスク [3] に合うように約 440 万件の

abstract を教師なし大規模コーパスとして使用した。こちらは以下で述べる教師なし学習にて使用した。

4.2 サブワードの学習

サブワードの学習には Sentence Piece[6] を用いた。SentencePiece とは決められた語彙数の辞書を教師なしコーパスから作成し、その辞書に沿った分割をすることができるトークナイザである。その手法として、SentencePiece では Byte Pair Encodeing やユニグラム言語モデル [5] が実装されている。今回の実験ではユニグラム言語モデルを使用し学習を行った。データセットには 4.1 節で述べた PubMed コーパスを用いて、複数の系列を得るために語彙数が 2,000、4,000、8,000 となるように辞書を学習した。

4.3 言語モデルの事前学習

Contextual String Embeddings と Contextual Subword Embeddings はそれぞれ以下のように学習した。

・***Contextual String Embeddings***

Contextual String Embeddings が実装されているフレームワーク [2] では 2015 年までの PubMed abstract の 5%を使用して学習されたモデルが配布されており、本研究ではそれを使用した。このモデルは 1,150 次元の LSTM 層を 3 層積み重ねた言語モデルである。

・***Contextual Subword Embeddings***

Contextual Subword Embeddings の学習データには 4.1 節の PubMed の大規模コーパスを使用した。このモデルでも 1,150 次元の LSTM を 3 層積み重ねた言語モデルを使用した。このモデルの学習は Contextual String Embeddings[1] 内の設定を参考とした。学習率ははじめに 20.0 とし、10 エポックに改善がなかった場合 1/4 に学習率を減らしていき、これを数回繰り返し学習させた。

4.4 CHEMDNER タスクの学習

今回の学習では複数サブワードを用いた BiLSTM-CRF モデルの有効性を調査するために複数のモデルで実験を行った。ベースラインは Akbik[1] らのモデルをベースラインとする。

提案モデルは上のベースラインモデルに加えて、4.3 節で事前学習させた Contextual Subword Embeddings を用いる。実験結果では、ベースラインに 2,000 の語彙数で分割したサブワードを接続した場合は +SW2k と表し、ベースラインに 2,000、4,000、8,000 を全て並列に接続した場合は+SW2k,SW4k,SW8k と表す。

表 1: CHEMDNER タスクによる実験結果

	Precision	Recall	F 値
BaseLine	91.05%	90.99%	91.02%
+SW2k	91.05%	91.60%	91.20%
+SW4k	90.79%	91.84%	91.31%
+SW8k	89.97%	92.15%	91.04%
+SW2k.4k	90.46%	91.85%	91.15%
+SW2k.8k	90.83%	91.27%	91.05%
+SW4k.8k	91.34%	91.55%	91.45%
+SW2k.4k.8k	91.19%	91.68%	91.43%

化合物名抽出の学習では検証データを用いて最適なパラメータを求めた。ハイパーパラメータ探索にはランダムサーチを使用した。パラメータはドロップアウトを {0.3, 0.4, 0.5}、単語レベルの LSTM の次元数を {100, 200, 300, 400}、単語レベルの LSTM の層数を {1, 2}、バッチサイズを {8, 16, 32, 64}、学習率を {0.025, 0.05, 0.075} の中から選択し、50 エポックを 10 通りの組み合わせで学習して最も高い F 値のパラメータを求めた。求めたハイパーパラメータで再度 150 エポックの学習を行った。また、学習率は 5 エポック中に改善が見られなかった場合、学習率を半減させる annealing 法を使用している。

4.5 実験結果

実験結果を表 1 に示す。Contextual Subword Embeddings を加えると全てのモデルで抽出性能が向上した。特に Contextual Subword Embeddings を複数加えている+SW2k.4k と SW2k.4k.8k では約 0.4pt 向上した。Contextual Subword Embeddings を加えたモデルでは Precision よりも Recall の増加が見られた。Precision はベースラインよりも低いものがあるが、全体として F 値が高い値となっているので、Contextual Subword Embeddings はテキストからの化合物名抽出に有効であると考えられる。

また、サブワード情報を加えることで未知語に対する抽出性能が向上すると考え、未知語のエンティティのみの実験を行った。未知語エンティティは、訓練データと検証データに出現しておらず評価データのみに出現する単語を含むエンティティとした。未知語エンティティは全部で 7,483 語で、そのユニーク数は 4,135 語であった。

未知語エンティティに対しての実験結果を表 2 に示す。未知語に対しての実験結果も、全体での実験結果と同様に Contextual Subword Embeddings を加えたモデルの方が良い抽出性能となった。特に

表 2: 未知語エンティティに対しての実験結果

	Precision	Recall	F 値
BaseLine	85.26%	85.43%	85.34%
+SW2k	86.08%	87.22%	86.64%
+SW4k	85.41%	87.56%	86.47%
+SW8k	84.30%	88.74%	86.64%
+SW2k.4k	85.47%	88.11%	86.77%
+SW2k.8k	85.72%	86.87%	86.29%
+SW4k.8k	86.30%	87.51%	86.90%
+SW2k.4k.8k	86.40%	88.08%	87.23%

図 2: ethylenedioxythiophene を複数のサブワード分割法で区切った結果

Contextual Subword Embeddings を複数加えている +SW2k.4k.8k では F 値は約 2pt 向上した。その他のモデルでも、全体の単語に対する結果より未知語に対する結果の方が F 値の上昇幅が大きくなっていることが分かる。

次に、ベースラインモデルでは抽出できなかったが、Contextual Subword Embeddings を加えることによって抽出できたエンティティを確認する。

例えば、”glyphosate isopropylamine”は、SW4k では ”g_ly_pho_s_ate iso_propyl_amine ”と分割される。この分割により “isopropylamine”という化合物名は、化合物が直鎖で繋がることを示す化合物の接頭辞である “iso”、プロピル基を表す “propyl”、アミンを表す “amine”で構成されていることが分かる。このように “isopropylamine”は未知語であるが化合物名に適した分割をすることで、新たに化合物名として抽出することが可能となったと考えられる。複数サブワード系列を加えることが効果的だったと考えられる例もあった。“poly(3,4-ethylenedioxythiophene)”では、SW2k と SW8k の分割方法によって得られる情報が変わってくる。SW2k ではエチル基を表す “ethyl”が獲得でき、SW8k ではエチレンを表す “ethylene”や、チオフェンを表す “thiophene”などを獲得できている (図 2)。サブワード分割の粒度により取れる化合物に関する名前は変わるため、複数のサブワード分割による情報を含むモデルの方が抽出性能が向上すると考えられる。

5 おわりに

本研究では複数のサブワード情報を BiLSTM-CRF モデルに加えることで、テキストからの化合物名の抽出性能が向上し、最も良いモデルでは F 値が 91.45% であった。特に未知語に対しては、サブワード情報を加えることで抽出できる化合物名が増加し、サブワードを用いないモデルよりも F 値が 2pt 向上した。

今回は複数の粒度のサブワードを加えたモデルを使用したが、どの粒度のサブワード情報が有効かは定かでは無い。サブワードを区切る粒度によって性能が変化してしまうので、粒度を決定的に決めるのではなく、統計的に求められるようなモデルが必要であり、今後検討したい。

参考文献

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

[2] flairNLP. flairnlp/flair: A very simple framework for state-of-the-art natural language processing (nlp), 2018. `https://github.com/zalandoresearch/flair`.

[3] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, Vol. 7, No. S1, jan 2015.

[4] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel Lowe, Roger Sayle, Riza Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, p. S2, 03 2015.

[5] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, 2018.

[6] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, 2018.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[8] Miloslav Nic, Jiri Jirat, and Bedrich Kosata. IUPAC compendium of chemical terminology (gold book), online version, 2012.