無機化合物を対象とした論文に対する化学物質名抽出システムの性能分析

町 光二郎 [1] 吉岡 真治 [2]

[1] 北海道大学工学部
[2] 北海道大学大学院情報科学研究院, 大学院情報科学院, 国際連携研究教育局 GSB, 創成研究機構化学反応創成研究拠点（WPI-ICReDD)
machi@eis.hokudai.ac.jp, yoshioka@ist.hokudai.ac.jp

1 はじめに

近年、論文からの知識発見の研究が盛んとなり、化学分野においても、論文からの化学物質に関する記述を抽出する化学物質名抽出（Chemical Named Entity Recognition: CNER）システムの研究が行われている。この研究の初期段階では、辞書やパターンをベースにしたシステム [1] の開発もあったが、近年は、コーパスを用いた機械学習によるシステムが、非常に高い認識性能を持つ様になってきている [2]。しかし、このコーパスの作成には、多大なコストがかかるため、大規模なものは限られた数しか存在しないだけでなく、その分野が生命医化学の分野に限られている [3, 4]。そのため、これらのコーパスからの機械学習をベースとしたシステム [5] を用いて、無機化合物を扱う分野を対象とした論文に対する化学物質名抽出を行う場合に、コーパスに存在しない無機化合物に関する再現率が低いという問題があった [6]。これに対し、近年のニューラル言語モデルをベースにしたシステムとサブワードによる単語分解の枠組が導入されることにより、システマティックに記述されることの多い化学物質については、コーパスに存在しなくても、抽出可能なケースが増えているのではないかと考えた。

本研究では、生命医化学の分野で学習したニューラル言語モデル CNER システムを用いて、無機化合物を扱った論文に対して抽出実験を行い、結果の分析を行った上で、その有用性について検討する。

2 既存の化学物質名抽出システム

2.1 予備的検討

我々は、これまでにナノ結晶デバイス開発の分野を対象として、論文からの実験情報の抽出を目標としたコーパスの作成 [7] や、そのコーパスを用いた自動情報抽出システムの構築 [6] を行なってきた。[6] では、当時の研究として、認識性能が高いと考えられていたコーパスからの機械学習を用いた ChemSpot [5] を CNER システムとして使用したが、コーパスに存在しない無機材料などの再現率が低いという問題が指摘されていた。

2.2 ニューラル言語モデル

近年、固有表現抽出の分野では、Bidirectional-Long Short-Term Memory (BiLSTM) と CRF を組み合わせた BiLSTM-CRF [8] や、双方向 Tranformer を使用した汎用言語モデルの BERT [9] などの、ニューラル言語モデルを利用したものが高い成果を挙げている。

ニューラル言語モデルの利点として、次のようなことが考えられる。

- 入力を時系列データとして処理することで、コンテキストを考慮した抽出が期待できる。
- 最小単位として単語やサブワード、文字を使用することで、抽出したい固有表現によく見られる特徴を捉え、コーパスに含まれていない固有表現や、記法が異なるものの抽出が期待できる。

また、事前学習時に大量の文書データを用いることにより、用意したコーパスに含まれないような表現でも、ある程度対応することが期待できる。

本研究では、BERT の一種である BioBERT [2] を抽出システムとして使用した。

2.3 BioBERT

BioBERT は、BERT を一般的な文（Wikipedia）で事前学習した後、追加で、生命医化学の分野の文章（PubMed+PMC）で再度事前学習を行なったものである。入力のためのサブワードは、WordPiece [10] が用いられている。図 1 のように、事前学習後に Fine-tuning をすることで、様々なタスクに応用できる。

固有表現抽出を行う場合は、一文ごとに WordPiece でトークンに分割し、それぞれのトークンについて、IOB ラベルの予測を行う。

3 無機化合物を対象とした抽出実験

3.1 目的

2.1 節に記述したように、[6] では、主に訓練用のコーパス不足のため、無機化合物の化学物質名に関する再現率が低かった。これに対し、2.3 節で紹介した BioBERT では、無機化合物に関する多くの Wikipedia 記事を事前学習に利用するといった対策が行われており、無機化合物に対しての認識性能の向上が期待される。

そこで、本研究では、生命医化学の分野のコーパスで学習した最先端のモデルを、無機化合物を扱う分野に対して利用し、どの程度機能するかを調べ、今後の利用可能性について検討する。

3.2 評価データの作成

評価データには、[7] のナノ結晶デバイス開発分野の論文 5 件を用いた。このコーパスは、化学物質名（SMaterial) や物質の特性（MChar)、実験パラメータ（ExP）などについてアノテーションされており、今回は、化学物質名にあたる SMaterial を対象とした。しかし、SMaterial の中には、族に関する表記（「III」や「V」など）や、固有表現の境界を間違えているもの、明らかに化学物質名ではないものが一部含まれていたので、評価データから除去した。

その結果、表 1 のようなデータを得られた。

表 1: 各論文における化学物質名の数

論文	1	2	3	4	5	total
延べ数	126	101	62	180	164	636
異なり数	17	17	13	23	18	40

3.3 実験

事前学習済みの BioBERT[1] を、生命医学の論文アブストラクトからなるコーパス CHEMDNER [3] を用いて学習を行ない、評価データに対して化学物質名の抽出を行なった。BioBERT では、固有表現抽出を行うために、コーパスを CoNLL フォーマットにする必要があるが、学習には、変換済みのデータ[2]を使用し、評価データは、既存の手法[3]で変換を行なった。

3.4 結果と考察

抽出結果は、単語中の部分文字列として与えられるため、正解とする化学物質名との比較においては、語の境界まで完全に一致している完全マッチと、語の境界については誤っているが化学物質名の記述が存在することは認識可能な部分マッチという二つのレベルの判定基準を用いて評価を行う。表 2 に、これらの二つの基準で評価を行った結果についての精度、再現率、F 値を示す。また、無機化合物名の記述スタイルにより、抽出性能が異なる可能性があるかどうかを検討するために、異なり数の再現率についても、表 3 にまとめた。

表 2 を見ると、再現率にやや問題があるようにも判断できるが、表 3 を見ると、コンテキストに依存して抽出に失敗している場合があるものの、論文中の一部では、化学物質名として抽出されていることが確認された。このような再現率の向上は、3.1 節で議論したことによるものであると考えられる。

表 2: 抽出システムによる抽出結果

	完全マッチ			部分マッチ		
論文	精度	再現率	F 値	精度	再現率	F 値
1	0.82	0.89	0.85	0.91	0.98	0.95
2	0.76	0.82	0.79	0.90	0.97	0.93
3	0.81	0.92	0.86	0.89	1.00	0.94
4	0.89	0.96	0.92	0.93	1.00	0.96
5	0.87	0.95	0.91	0.92	1.00	0.96
全体	0.84	0.92	0.88	0.91	0.99	0.95

抽出結果と正解データを比較すると、次の 2 つの特徴が見られた。

[1] https://github.com/dmis-lab/biobert
[2] https://github.com/cambridgeltl/MTL-Bioinformatics-2016
[3] https://github.com/spyysalo/standoff2conll

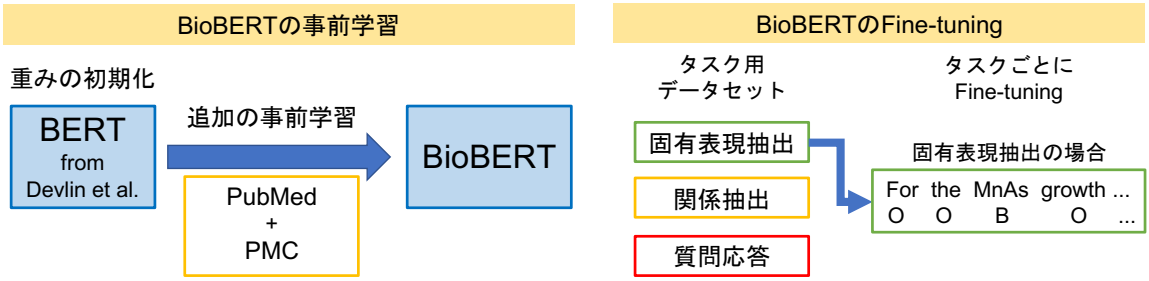

図 1: BioBERT の概観

表 3: 各論文における異なり数の再現率

文書	1	2	3	4	5	全体
完全	0.94	0.94	0.92	1.00	0.94	0.97
部分	1.00	1.00	1.00	1.00	1.00	1.00

- 分野特有の記法に対して、境界を間違えることがある（図 2）。
- 論文全体を見れば容易に判別できるものを間違えることがある（図 3、図 4）。

予測: p[(CH3C5H4)2Mn], InP (111)B
正解: p[(CH3C5H4)2Mn], InP (111)B

図 2: 分野特有の表記に関する誤り

図 2 の「p[(CH3C5H4)2Mn]」や「InP (111)B」は、実験のパラメータや、物質の特性を表す記法である。生命医化学とは異なる分野において用いられるが、生命医化学の分野であまり現れない記法については、学習データが不足しているために、化学物質名の語の境界を誤ることが多いことが確認された。

... by the ones under the external magnetic fields H applied at θ and ... .
... in which H was always applied in ... .

図 3: 論文 2 における「H」の判別

図 3 の「H」は、論文 2 の中で、磁場の強さという意味で用いられている。最初に書かれた文では抽出されていないが、それ以降は、抽出されてしまっていた。これは、BioBERT が文ごとに抽出を行なっており、論文全体の情報を考慮できないためであると考えられる。

図 4 の「V」は、バナジウムとも V 族という意味とも取れるが、これも、論文全体を見れば、判別がつくものである。もしくは、この分野において、多くの場合、「V」が V 族という意味で使われるという知識があれば間違えにくいとも考えられる。

... diluted in H2 were used as group III and V source materials.
The Tg and the V/Mn ratio were 650 ° C and 1125, respectively.

図 4: 論文 2 における「V」の判別

3.5 無機化合物名抽出における有用性

今回の実験で、論文 5 件という少ない量のコーパスではあるが、ニューラル言語モデルを用いた最新の CNER システムは、従来の生命医化学の分野のコーパスのみを利用していたシステムに比べて、無機化合物を対象とした論文からも、非常に高い性能で、化学物質名の抽出ができることが確認された。しかし、分野特有の記法などを考慮した語の境界の認定や、論文全体を考慮した一貫した化学物質名の抽出という観点からの問題があることも確認された。今後は、これらの問題に対して、論文全体を考慮することで、一貫した抽出結果を得るようにする後処理について検討をしていきたい。

具体的には、

- 分野に関する知識をもとに、いくつかのパターンによって判別する。
- 3 値分類「化学物質名である・化学物質名ではない・どちらの場合もある」をする学習器を作成し、どちらの場合もあるものについては、その都度人手で判別する。

このような処理を加えることにより、主に有機化合物を扱うコーパスで学習した抽出システムでも、無機

化合物を扱う分野で利用することは、十分期待できると考えられる。

また、この様に作成した抽出結果を含むデータを、論文データベース [11] として、実際に分野の専門家に提供するとともに、確認をしてもらうことで、分野に特化したコーパスの作成についても検討していきたい。

4 おわりに

ここまで、主に有機化合物を扱うコーパスで学習した化学物質名の抽出システムが、主に無機化合物を扱う分野において、どの程度機能するかについて実験を行なった。その結果、論文 5 件という少ない数ではあるが、十分機能することが確認できた。

今後は、後処理の方法を検討するとともに、分野の専門家に論文データベースを提供し、その有用性を示すことで、抽出結果を確認してもらう予定である。

謝辞

本研究の一部は、JSPS 科研費 19K22888 の助成と北海道大学創成研究機構化学反応創成研究拠点（ICReDD）の支援を受けた。ここに記して謝意をあらわす。

参考文献

[1] David M. Jessop, Sam E. Adams, Egon L. Willighagen, et al. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminformatics*, Vol. 3, p. 41, 2011.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019.

[3] Martin Krallinger, Obdulia Rabal, Florian Leitner, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminformatics*, Vol. 7, No. S-1, p. S2, 2015.

[4] Corinna r, Roman Klinger, Christoph Friedrich, et al. Chemical names: Terminological resources and corpora annotation. pp. 51–58, 01 2008.

[5] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, Vol. 28, No. 12, pp. 1633–1640, 2012.

[6] Thaer M. Dieb and Masaharu Yoshioka. Extraction of chemical and drug named entities by ensemble learning using chemical NER tools based on different extraction guidelines. *Trans. MLDM*, Vol. 8, No. 2, pp. 61–76, 2015.

[7] Thaer M. Dieb, Masaharu Yoshioka, and Shinjiroh Hara. Nadev: An annotated corpus to support information extraction from research papers on nanocrystal devices. *JIP*, Vol. 24, No. 3, pp. 554–564, 2016.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, Vol. abs/1508.01991, , 2015.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[10] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, Vol. abs/1609.08144, , 2016.

[11] 吉岡真治, 尹磊, 原真二郎ほか. 専門用語の知識保全エコシステムを有するインハウス論文・図表データベースの構築. 言語処理学会第 25 回年次大会発表論文集, B4-7, 2019.