

自己学習による化学文書中の専門用語抽出

Chemical Named Entity Recognition with Self-Training

崔一鳴 ^{*1*3}

Yiming Cui

西川仁 ^{*1*3}

Hitoshi Nishikawa

徳永健伸 ^{*1}

Takenobu Tokunaga

吉川和 ^{*2*3}

Hiyori Yoshikawa

岩倉友哉 ^{*2*3}

Tomoya Iwakura

^{*1}東京工業大学 情報理工学院

School of Computing, Tokyo Institute of Technology

^{*2}株式会社富士通研究所

Fujitsu Laboratories Ltd.

^{*3}理研 AIP-富士通連携センター

RIKEN AIP-Fujitsu Collaboration Center

In this paper, we propose to use self-training for chemical named entity recognition. We first train a neural network-based model for chemical named entity recognition model using the CHEMDNER corpus. The trained model is used to annotate the unlabelled MEDLINE corpus to create automatically labelled training data. We then use both training data, manually labelled CHEMDNER corpus and automatically labelled MEDLINE corpus, to train our final model. The evaluation using the unlabelled MEDLINE corpus as training data showed that the effectiveness of self-training in the chemical named entity recognition task.

1. はじめに

化学分野の研究は非常に盛んであり、日々新たな発見がなされ、論文が発表されている。それらの論文の中には今まで登場したことのない、新しい化学用語が出現する。化学用語のデータベースは化学分野の研究において重要な言語資源であるが、現状ではそのデータベースは新しく出版された論文や特許を人手によって読解し、新しく出現した用語を抽出し構築されている。この作業は費用を要する作業であるだけでなく、時間的な面においても困難であり、新しい用語の登場の速度に人手による抽出によって追従することは容易ではない。さらに、化学分野に精通した人員でなければ新しい化学用語の抽出は難しい作業であるため、人員の確保そのものも容易ではない。そのため、論文からの情報抽出の自動化は急務である。

この問題を解決するため、化学文書からの化学に関する専門用語の自動抽出を試みる研究が進められている [14]。化学用語の自動抽出課題のために作られたコーパスである CHEMDNER コーパス [4] においては、10,000 件の化学分野の論文のアブストラクト（訓練データ 3,500 件、開発データ 3,500 件、テストデータ 3,000 件）に対して、化学用語部分にラベルが付与されている。このコーパスにおいて現時点での最高精度を報告している研究は Bi-LSMT-CRF [5] に注意機構を加えたモデルであり、F 値 91.14% を達成している [6]。同じデータに対して、専門家が化学用語の抽出を行った結果の一致率は 89% であるため、それを上回る精度を達成したことになる。

本研究では、この化学用語抽出課題の精度をさらに向上させる試みとして、自己学習によって CHEMDNER 以外の大規模なデータを利用する手法を提案する。具体的には、CHEMDNER の訓練データを用いて作成したモデルを用いて、化学用語部分にラベルが付与されていない MEDLINE コーパス [11] のアブストラクトにラベルを付与し、それを新しい訓練データとして、化学用語抽出のモデルを作成する。

2. 関連研究

化学用語の抽出については数多くの研究がなされており、類似する課題である固有表現抽出 [2] と同様に、機械学習を用いる手法が数多く提案されている。近年は、ニューラルネットワークを用いる手法が活発に研究されており、CHEMDNER コーパスにおける現時点での最高の性能を報告している論文 [6] は双方向 LSTM (Bi-LSMT) と条件付き確率場 (CRF) を組み合わせた Bi-LSMT-CRF [3, 7] に基づく。先行研究に倣い、本研究でも Bi-LSMT-CRF をベースラインとして学習を行い、モデルを作成する。Luo らは Bi-LSMT-CRF に注意機構 [1] を加えたモデルを採用し、最高精度を報告している。Luo らは入力の特徴量として、MEDLINE コーパスと CHEMDNER コーパスで word2vec を用いて単語分散表現を用いている。また、品詞データや Bi-LSMT による文字分散表現 [12] も入力の特徴量として用いている。

自己学習は自然言語処理において広く用いられており、構文解析 [8] や語義曖昧性解消 [9] などに利用されている。

ニューラルネットを利用したモデルに自己学習を利用した先行研究として竹前らが見出し生成課題に自己学習を利用したものの [16] がある。竹前らは、正例が付与されているデータからまずモデルを作成し、それを正例が付与されていないデータに対して適用することで疑似的な正例データを作成した。その上で正例が付与されているデータと、疑似的な正例データの両者を用いて最終的なモデルを作成した。自己学習を行ったモデルと行っていないモデルを比較することによって、竹前らは見出し生成課題において自己学習を行うことによって性能が向上することを示した。

本論文は、化学用語抽出課題において、自己学習を行うことを提案する。まず化学用語ラベルが付与されているデータセットである CHEMDNER コーパスを用いてモデルを学習し、化学用語ラベルが付与されていない MEDLINE コーパスに疑似的な正解ラベルを付与する。その後、CHEMDNER コーパスと疑似的な正解ラベルが付与された MEDLINE コーパスのデータを併せて再度モデルを作成し、これを最終的なモデルとする。

連絡先:

^{*1}{sai.m.ab@m, hitoshi@c, take@c}.titech.ac.jp^{*2}{y.hiyori, iwakura.tomoya}@jp.fujitsu.com

3. 提案手法

本論文の提案手法は以下のような手順になる。

1. 教師あり学習：まず、CHEMDNER の訓練データを利用して教師あり学習を行う。これをベースラインモデルとする。
2. 疑似教師データの作成：次に化学用語のラベルが付与されていない、テキストのみのデータに対して手順 1 で作成したモデルを用いて、化学用語ラベルを付与する。これを疑似訓練データとする。
3. 新規モデルの学習：手順 1 で利用した CHEMDNER の訓練データと、手順 2 で化学用語ラベルを付与した疑似訓練データの両方を用いて、新しいモデルの学習を行う。その後、CHEMDNER のテストデータを用いて、このモデルの精度の評価を行う。
4. 手順 2 と手順 3 を繰り返し、最終的なモデルを得る：手順 3 で得られたモデルの性能の評価が手順 1 で得られたモデルよりも高精度である場合、手順 3 で得られたモデルを再度用いて手順 2 と手順 3 を行う。精度向上がみられなかった場合学習を終了し、最終的なモデルを得る。

4. 実験

4.1 ベースラインモデル

ベースラインのモデルは、単語分散表現と Bi-LSMT-CRF を利用している。構成図を図 1 に示す。

4.1.1 単語分散表現

単語系列をモデルに入力する際には、word embeddings 層を通して、単語を単語分散表現に変換している。その際には gensim^{*1} による word2vec[10] を利用した。使用データは CHEMDNER コーパスの訓練データおよび単語分散表現を得る際には、スペースで区切られているものを 1 単語とし、パラメータとして次元数は 100、window size は 4、iter は 10 とした。また min count は 0 とし、学習時に登場した単語を全て分散表現の辞書に登録した。

加えて、文字の分散表現の情報も gensim ライブラリの word2vec を用いて獲得した。文字分散表現は 10 次元とし、こちらも CHEMDNER コーパスの訓練データおよび MEDLINE コーパスを利用した。

最終的には、ある単語の単語分散表現は、単語そのものの分散表現 100 次元と、単語を構成する文字の分散表現 10 次元の合計 110 次元となる。文字の分散表現は、単語に含まれている文字分散表現の平均を取った 10 次元とした。

4.1.2 Bi-LSMT-CRF

本研究におけるベースラインのモデルは Bi-LSMT-CRF によって構築されている。これは PyTorch^{*2} を用いて実装を行った。LSTM の隠れ層は 150 次元とした。また、テストの際に、分散表現辞書にない未知語が現れた際には、ランダムなベクトルを生成し、割り当てた。登場する単語が未知でも、その単語に含まれている文字情報は未知ではないことが多い。単語が未知の場合は、ランダムに生成した 100 次元のベクトルと、文字情報から得られるランダムではない 10 次元のベクトルを合成して、110 次元のベクトルとした。また、本研究で

は未知語にランダムなベクトルを割り当てる際に、同じベクトルが割り当てられるよう seed 値を固定して乱数を生成した。

4.1.3 化学用語ラベル

化学用語ラベルには固有表現抽出課題で広く利用されている IOB2[13] タグ方式を採用した。これは、化学用語を構成する最初の単語に B、最初の単語ではないが化学用語の一部である単語には I、化学用語ではない単語には O というタグを付与するというものである。

4.1.4 学習

このモデルを CHEMDNER コーパスの訓練データを用いて学習を行ったものをベースラインのモデルとした。学習の際にはミニバッチ学習を行い、ミニバッチのサイズは 100 とした。また、epoch 数は最大 20 とした。

4.2 比較手法

- ベースライン手法：CHEMDNER コーパスの訓練データを用いて、前述のモデルを用いた化学用語抽出モデルを学習し、これをベースラインの手法として利用した。
- 提案手法：ベースラインの手法によって学習が行われたモデルを用いて、化学用語ラベルが付与されていない MEDLINE コーパスのデータに化学用語ラベルを付与し、それを疑似訓練データとし、前述したように CHEMDNER の訓練データと MEDLINE 疑似訓練データの両者を利用することで化学用語抽出モデルを構築する。

4.3 データ

4.3.1 CHEMDNER コーパス

CHEMDNER コーパスは化学用語抽出課題のために構築されたコーパスである [4]。コーパスの統計量を表 1 に示す。CHEMDNER コーパスは化学関連論文のアブストラクト 10,000 件からなり、化学用語の箇所に化学用語ラベルが付与されている。化学用語は、TRIVAL, SYSTMATIC, ABBREVIATION, FORMULA, FAMILY, IDENTIFIER, MULTIPLE, NO CLASS の 8 クラスに分類されラベルが付与されているが、本実験では先行研究 [6] と同様に、これら 8 クラスを等しく化学用語であるとして、すなわち単一のクラスとして扱う。

4.3.2 MEDLINE コーパス

MEDLINE は医学を中心とする文献情報を収集したオンラインデータベースである。[15] このデータベースは医学、薬学、看護学、歯学、衛生学、獣医学、生化学、分子生物学など医学に関連する幅広い文献情報を含んでいるが、本研究では、MEDLINE コーパスの中でも、2017 年版の CHEMDNER コーパス作成時に対象としたジャーナル・会議のもののみを利用した。MEDLINE コーパスには、CHEMDNER コーパスとは異なり、化学用語ラベルは付与されておらず、テキストのみのデータのみとなっている。使用した MEDLINE コーパスに含まれるアブストラクトの文字数の合計は約 18 億字であり、これは CHEMDNER コーパスの訓練データ約 488 万字の 375 倍の量に相当する。また、単語数は約 2 億 7 千万語あり、これは CHEMDNER コーパスの訓練データ約 77 万語の約 350 倍に相当する。前述したように、本研究ではこの MEDLINE コーパスに対して CHEMDNER コーパスで訓練したモデルを利用して化学用語ラベルの付与を行い、これを疑似訓練データとして利用する。

4.4 評価

学習を終えた化学用語抽出モデルは CHEMDNER コーパスのテストデータを利用して行う。また、本研究が対象とする

*1 <https://radimrehurek.com/gensim/>

*2 <https://pytorch.org/>

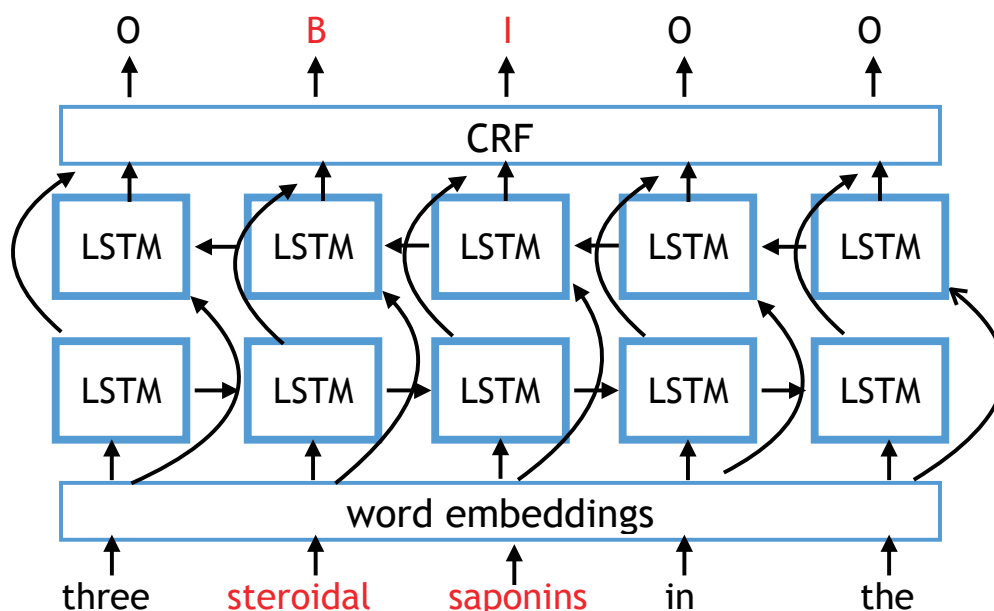


図 1: 実装したモデルの概念図

表 1: CHEMDNER コーパスの基本データ

	訓練データ	開発データ	テストデータ	合計
アブストラクト数	3,500	3,500	3,000	10,000
全文字数	4,883,753	4,864,558	4,199,068	13,947,379
全単語数	770,855	766,331	662,571	2,199,757
TRIVIAL	8,832	8,970	7,808	25,610
SYSTEMATIC	6,656	6,816	5,666	19,138
ABBREVIATION	4,538	4,521	4,059	13,118
FORMULA	4,448	4,137	3,443	12,028
FAMILY	4,090	4,223	3,622	11,935
IDENTIFIER	672	639	513	1,824
MULTIPLE	202	188	199	589
NO CLASS	40	32	41	113

表 2: 実験結果

使用データ	精度	再現率	F 値
ベースライン	0.866	0.787	0.824
自己学習 1 回	0.867	0.812	0.839
自己学習 2 回	0.857	0.826	0.841
自己学習 3 回	0.842	0.837	0.839

化学用語抽出課題と同様に線形ラベル付け問題である固有表現抽出課題においては、一般的に精度、再現率、および F 値が評価尺度として用いられるため、本研究においてもこれらの値を評価尺度として用いる。

5. 結果と考察

CHEMDNER コーパスのテストデータを用いて評価した結果を表 2 に示す。ベースラインモデルに比べて、自己学習を行うことで F 値が向上していることがわかる。特に、提案手法の手順を繰り返すことによって再現率が繰り返し向上しており、このことは訓練データの網羅性の不足が自己学習によって

補われていることを示唆している。一方で、精度は自己学習を 1 回だけ行った場合が最もよい。これは自己学習によって本来は正しくない単語列が化学用語として疑似訓練データに混入することが増えることによって、テストデータにおける精度が低下するためと思われる。結果としては自己学習を 2 回行った際の F 値が最も良好な結果を得た。

また、手順 2, 3 で用いるデータの量を変化させた際の精度の変化を表したグラフを表 3 に示す。この結果は自己学習を 1 回だけ行った際の結果である。追加データの量は、CHEMDNER の訓練データの量を 100% とした際の疑似訓練データの量を示す。すなわち、0% は自己学習を行わない場合（表 2 のベースライン）に相当する。全体的に、疑似訓練データを増やすことで、ベースラインのモデルと比較して F 値が向上することがわかる。結果として 450% において最高精度を記録しているものの、その一方で、データ量を増加させたからといって一貫して F 値が向上するという傾向は見られず、データを追加したからといって安定して性能が向上するとは言えない。そのため、自己学習に利用するデータを増加させることによって安定して性能を向上させるためには何らかの追加的な工夫が必要となるものと思われる。

表 3: データ量ごとの結果

追加データの量	精度	再現率	F 値
0 %	0.866	0.787	0.824
50 %	0.879	0.789	0.832
100 %	0.870	0.808	0.838
150 %	0.860	0.812	0.838
200 %	0.867	0.812	0.839
250 %	0.877	0.800	0.836
300 %	0.849	0.832	0.841
350 %	0.877	0.800	0.836
400 %	0.875	0.801	0.836
450 %	0.883	0.804	0.842
500 %	0.860	0.822	0.841
550 %	0.883	0.804	0.837

6. まとめ

本論文では、Bi-LSMT-CRF による化学用語抽出のモデルを学習する際に、自己学習を利用してラベルの付与されていないデータを利用することを提案した。実験により、自己学習を利用することでモデルの性能が向上したことが示された。また、自己学習を利用して得られたモデルで再度ラベルの付与を行い、それを利用して再び自己学習を行うことで、さらに性能が向上することも確認した。自己学習の際に利用するデータを増加させることによって F 値が向上することも確認できたものの、F 値の向上は一定しておらず、データ量を増加させる際には何らかの工夫が必要であると思われる。本研究は最高精度に到達していないが、現在最高精度を出しているモデルを再現し自己学習を加えることで、最高精度に到達できる見込みがあるものと思われる。

参考文献

- [1] Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1462–1472, 2016.
- [2] Asif Ekbal and Sivaji Bandyopadhyay. Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical, Computer, and Systems Engineering*, Vol. 4, No. 2, pp. 155–170, 2010.
- [3] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
- [4] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, Roger A Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktschel, Sergio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko itnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A Akhondi, Jan A Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usi, Rui Alves, Isabel Segura-Bedmar, Paloma Martnez, Julen Oyarzabal, and Alfonso Valencia.

The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, No. Suppl 1, pp. 1–17, 2015.

- [5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.
- [6] Ling Luo, Zhihao Yang, Pei Yang, Zhang Yin, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2018.
- [7] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.
- [8] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pp. 152–159. Association for Computational Linguistics, 2006.
- [9] Rada Mihalcea. Co-training and self-training for word sense disambiguation. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, 2004.
- [10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [11] U.S. National Library of Medicine. Medline. https://www.nlm.nih.gov/databases/download/pubmed_medline.html.
- [12] Marek Rei, Gamal KO Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*, 2016.
- [13] Erik F Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 173–179. Association for Computational Linguistics, 1999.
- [14] Miguel Vazquez, Martin Krallinger, Florian Leitner, and Alfonso Valencia. Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics*, Vol. 30, No. 6-7, pp. 506–519, 2011.
- [15] Beatriz Vincent, Maurice Vincent, and Carlos Gil Ferreira. Making pubmed searching simple: learning to retrieve medical literature through interactive problem solving. *The oncologist*, Vol. 11, No. 3, pp. 243–251, 2006.
- [16] 竹前慎太郎, 村尾一真, 谷塚太一, 小林隼人, 野口正樹, 西川仁, 徳永健伸. 自己学習を用いたニューラル見出し生成. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 3Pin136–3Pin136. 一般社団法人 人工知能学会, 2018.