Vol.2020-IFAT-139 No.3
2020/7/31

化学物質データベースを訓練データに用いた化学物質名識別システムに関する実験的分析

町 光二郎[1,a)]　吉岡 真治[1,b)]

概要：機械学習による論文からの化学物質名抽出は，一般に人手によるアノテーションが行われたコーパスを用いた系列ラベリングのタスクとして行われる．しかし，人手によるコーパスの作成が高コストであるため，コーパス中に含まれる化学物質名のバリエーションは，限定されたものとなる．一方，化学物質名は，ある種の命名規則によって名づけられることが多いため，単語中の文字列に注目した機械学習により，化学物質名の判定を行う枠組みを考えた．本研究では，化学物質に関する大規模なデータベースである ChemIDplus の化学物質名と機械可読辞書から得られた名詞を訓練データとして使う化学物質名識別システムを提案する．本稿では，本システムの性能を負例である名詞の選定方法とトークンの分割法の与える影響に注目して分析を行った結果について報告する．

Experimental Analysis of Chemical Named Entity Recognition System Using Chemical Database for Training Data

Abstract: Machine learning based Chemical Named Entity Recognition (CNER) task can be formalized as a sequence labeling task that uses human annotated corpus as training data. However, due to the cost of constructing such corpus, the variation of the chemical named entities are limited. On the contrary, since names of chemical named entities are generated by using naming conventions, it is possible to construct CNER system that classify words into chemical named entity or not based on the analysis of their character sequences. In this paper, we propose such a CNER system using ChemIDplus that covers wide varieties of chemical named entities and noun words from machine readable dictionary as a training data. In this paper, we also analyze the effect of the method for constructing non chemical words list and one for splitting words into tokens using experimental analysis of the system.

1. はじめに

近年，固有表現抽出の分野では，BiLSTM-CRF [1], [2] や BERT [3] など，深層学習による文脈を考慮した研究が成果を挙げており，化学物質名抽出（Chemical Named Entity Recognition， CNER）の分野でも，それらの研究を分野適応させたもの [4] が開発されている．また，近年のモデルでは，単語中の特徴的な文字列であるサブワード [5] や文字をトークンとして使用することで，字面の似ている単語の特徴を捉えた手法が成果を上げている [4], [6]. 一方，一般に化学物質名はシステマティックな記述が行われることが期待されるが，非常に多くの物質が存在するため，その多くをカバーするようなコーパスを作成することが困難である．そのため，コーパス中に似たような化学物質名が存在しない場合に，CNER システムが抽出に失敗するといった問題がある．

その他にも，機械学習を用いた抽出システムは，文単位で抽出を行うものが多いため，文書内の同じ単語を化学物質として認識する場合としない場合があるという問題が起きる可能性がある．この問題に対し，ルールベースで後処理を行う方法 [7] や，文書全体から予測を行うモデル [8] が提案されている．しかし，前者は，機械学習モデルが誤抽出した語をさらに抽出してしまう可能性があり，後者は，それ以降に開発された文を単位として抽出を行う別のモデル [4] の方がスコアが良いという結果となっている．

我々は，これらの問題に対し，化学物質名データベースを用いることで得られる大規模な訓練データを用いて，単

[1] 北海道大学
〒 064-0806 札幌市北区北 14 条西 9 丁目
[a)] machi@eis.hokudai.ac.jp
[b)] yoshioka@ist.hokudai.ac.jp

語を単位とした識別システムを提案する．一般に，多くの化学物質名は，命名規則などに基づいて作成されることが多いことから，これらの規則を抽出するためのサブワードもしくは文字列を利用した LSTM による系列分析を行う方法を提案する．

本稿では，主に学習データの選定とトークンの分割について議論するために，化学物質データベースと一般語の辞書から 2 つの訓練データを作成し，複数のトークン分割による学習と評価を行った．さらに，訓練データから学習した特徴の汎用性を確認するために，論文中にに出現する化学物質名に対してモデルを適用した上で，訓練データとトークン分割の観点から分析を行った．

2. LSTM を用いた識別システム

Long Short Term Memory（LSTM）[9] は，入力を再帰的に処理する Reccurent Neural Network（RNN）を改良したものであり，時系列データの処理によく用いられる．自然言語処理の分野でも，文書分類や，固有表現抽出などの様々なタスクで用いられている．

本稿では，機械学習ライブラリの PyTorch [10] を用いて，1 層の忘却ゲート付き LSTM [11] による識別システムを作成した．分類は，LSTM の最終的な出力を 1 層の全結合層に入力することで行った．

入力の単位には，文字レベルと SentencePiece [12] を採用した．これらを比較することにより，語の特徴を捉えるためには，どの単位や語彙数が適切かについて議論する．

3. 訓練用データの準備

ここでは，訓練用のデータセットについて紹介するとともに，実験のための前処理について説明する．実験では，化学物質名のデータベースである ChemIDplus [13] を使用して正例のデータを作成し，一般語の概念辞書である WordNet [14] を使用して負例のデータを作成した．

3.1 ChemIDplus

ChemIDplus は，米国立医学図書館（National Library of Medicine）が提供する化学物質検索システムであり，約 42 万件の化学物質が登録されている．それぞれの化学物質は，**表 1** のように ID，化学式（displayFormula），見出し語（displayName）の他に，同義語や構造に関する情報を持つ．しかし，約 10 万件のものは，化学式が登録されていなかったり，その他にも「Unspecified」となっているものがある．本稿では，2020 年 3 月 28 日付の ChemIDplus から，化学式と見出し語を化学物質名のデータとして使用した．ただし，このデータベースには，論文中で用いられることがないと思われる記法や見出し語が存在する．そのため，次のような前処理を行い，その結果を訓練データの正例とした．

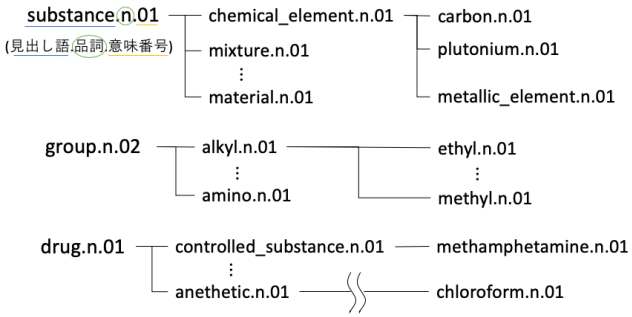

図 1 WordNet の階層関係
Fig. 1 Hierarchical relationship in WordNet

(i) 薬品の一般名に関する表記部分を除去
例：Dimethyltubocurarine[BAN]→Dimethyltubocurarine
(ii) 化学式が存在しないものと Unspecified のものは，見出し語も使用しない

（i）は，[BAN] などの表記が使用されにくいと思われるため，その記述部分を除去した．（ii）は，化学式が存在しないものと Unspecified となっているものについては，化学物質名の特徴を捉えるために不要だと思われる植物などの記述が多く存在したため，（例: Withania Somnifera - ナス科の低木），全てデータセットから除外した．ただし，Carotenoids のように化学物質名と思われるものも含まれていたため，全て除去すべきかどうかは，議論が必要である．

3.2 WordNet

WordNet は，米プリンストン大学で開発された概念辞書である．登録されている単語は，synset と呼ばれる意味を表す単位と対応づけられており，これによって同義語や多義語を表現することができる．単語が多義（複数の synset に対応する）の場合は，品詞と意味番号により，その単語に対応する synset を表す．また，各 synset は，階層関係によって結ばれており，上位語や下位語を辿ることで情報を得ることができる．**図 1** に，WordNet の階層関係の例を示す．ここでは，synset の ID の代わりに，先に述べた対応する単語，品詞，意味番号の組み合わせにより表現している．この図から分かるように，例えば，chemical element.n.01 の下位語を見ることで，carbon.n.01 や plutonium.n.01 などを取得することができ，ある概念に含まれる具体例を見ることができる．この階層には，多くの化学物質名が存在しており，負例として適さないものである．ただし，metallic_element のように化学物質名以外の語も含むので，全て除去すべきかどうかは議論が必要である．

本稿では，WordNet3.0 をデータとして使用し，処理を行う際には，Natural Language Toolkit [15] をツールとして利用した．訓練データは，全ての名詞を用いたデータと，以下に述べる処理によって一部の名詞を削除したデータに対して，それぞれ共通の前処理を加えることで負例として

表 1　ChemIDplus に含まれるデータの例
Table 1　Examples of Data in ChemIDplus.

ID	displayFormula	displayName
0000034742	C12-H14-O4	Monobutyl phthalate
0000035676	C39-H45-N2-O6	Dimethyltubocurarine [BAN]
0000036884	Unspecified	Carotenoids
ZC80600000	Unspecified	Withania Somnifera, extract
YY39705000	(Empty)	Vicoa indica Willd. D.C.

利用し，2 つのデータセットを作成した．一部の名詞を除去したデータは，先に述べた通り，下層に化学物質名を多く含むと思われる概念の下位語を全て除去することで作成した．具体的には，図 1 の substance.n.01 と group.n.02，drug.n.01 の 3 つに属する下位語を全て削除した．これらの語は，いくつかの化学物質名から上位語をたどることによって決定した．この処理を加えることで生じる問題については，先に述べた通りである．共通の前処理としては，下線符号による分割が使用されているため，通常の文と同様にスペース区切りに変換した（例: cobalt_blue→cobalt blue）．

4. データセットを用いた実験

ここでは，負例のデータの作成方法について議論するために，3.1 節で作成した訓練データの正例として，3.2 節で作成した 2 つのデータを負例とした実験を行い，両者について比較を行った．

4.1 実験設定

まず，3 節で用意したデータを**表 2** のように，訓練用 8 割，開発用 1 割，テスト用 1 割の割合で分割して 2 つのデータセットを作成し，2.2 節で紹介した LSTM による分類器を，文字ベース，語彙数 1000 と 4000 の SentencePiece（SP1k, SP4k）の 3 パターンでそれぞれ学習させた．学習に使用したパラメータは，埋め込み次元数 256，隠れ層の次元数 128，学習率 0.001，エポック数 5 としており，人手でこれらの調整を行った．損失関数には，交差エントロピー誤差を使用し，最適化には，Adam [16] ($\beta_1 = 0.9, \beta_2 = 0.999$) を用いた．

学習後，モデルをテスト用データに適用し，精度や再現率，F 値による評価とエラー分析を行った．

4.2 結果・考察

表 3 にそれぞれのデータセットにおける各モデルの予測結果を示す．どちらのデータセットでも ChemIDplus よりも WordNet の方がわずかに F 値が低いが，全てのモデルにおいて非常に良い値であったといえる．また，データセット間で F 値を比較すると，全てのモデルにおいて，負例を整備した方がわずかに良い結果となった．これは，

表 2　訓練データの概要
Table 2　Overview of training dataset.

	訓練用	開発用	テスト用
ChemIDplus(CID)	373044	46630	46631
WordNet(WN)	95227	11903	11904
WordNetchem(WNchem)*	89861	11233	11233
CID+WN	468271	58533	58535
CID+WNchem	462905	57863	57864

*WordNetchem は，WordNet から化学物質名を多く含むと思われる語の下位語を除去したもの

WordNet の化学物質名をある程度除去したことで，それらが評価時に誤抽出として扱われる数が減少したことや，訓練時に一般語の特徴として学習される数が減ったためであると考えられる．入力の分割法による違いは，あまり見られなかった．

次に，WordNet の名詞を全て使用したモデルについて，化学物質名が含まれていることによる影響について考える．これについては，学習時に WordNet に含まれる化学物質名を一般語の特徴として学習してしまったり，評価時に化学物質名として抽出したが，ラベルが WordNet となっているため誤りであるとされてしまうといったことが見られた．**表 4** は，WordNet を全て使用した場合の誤り例である．前者の例として，cyclodrine hydrochloride が一般語として抽出されてしまったのは，WordNet に cyclobenzaprine や hydrochloride などの cyclo-や hydro-が含まれる単語があるためであると考えられる．後者の例として，1-dodecanol や trifluoromethane は，化学物質名であるが，正解ラベルが WordNet であるため，誤分類とされてしまっている．負例を整備したデータでは，誤って ChemIDplus の語であると判定されたものの中に化学物質名が含まれていたことから，階層関係による除去にも漏れが生じていたことがわかった．一方で，誤って WordNet の語であると判定したものを見ると，全体を使用した場合と比較して，化学物質名の特徴的な記法を含む語が少ないように見えため，学習時における悪影響は緩和されていると考えられる．

5. 論文中に出現する化学物質名への適用

4 節で学習したモデルは，負例を整備した場合としない場合で誤り例の内容に違いが見られたが，データセット内の評価値だけを見ると，いずれも良い値であり差もわずか

表 3　異なる設定による訓練を行なった識別システムのテストデータに対する予測結果
Table 3　Test results using different training settings of recognition system

		CID+WN			CID+WNchem		
クラス	分割	精度	再現率	F 値	精度	再現率	F 値
ChemIDplus	char	0.997	0.995	0.996	0.996	0.998	0.997
	SP1k	0.996	0.997	0.997	0.998	0.997	0.997
	SP4k	0.997	0.996	0.996	0.997	0.997	0.997
WordNet	char	0.984	0.990	0.987	0.993	0.984	0.989
	SP1k	0.991	0.986	0.988	0.989	0.991	0.990
	SP4k	0.986	0.988	0.987	0.988	0.990	0.989

表 4　分類を誤った例
Table 4　Error examples.

正解ラベルが ChemIDplus	正解ラベルが WordNet
AMB	DDT
F	FET
beta methyl ionone	1-dodecanol
cyclodrine hydrochloride	Hylactophryne
Ytterbium phosphate	February 2

であった．しかし，実際の論文では，1 つの物質に対して様々な記法が存在するため，各データセットから学習した特徴が有効であると限らない．

この節では，実際の論文中に出現する化学物質をどの程度検出できるかを確認するために，4 章で学習したモデルを抽出タスク用のコーパスに対して適用する．

5.1 CHEMDNER

CHEMDNER [17] は，PubMed に登録されている論文アブストラクト 10000 件について，化学物質名のアノテーションがされているコーパスである．このコーパスは，化学物質名が SYSTEMATIC や ABBREVIATION などの 7 クラスでアノテーションされており，クラスごとに評価を行うことで，識別システムの特徴を分析するのに役立つとが期待される．

前処理は，訓練データと開発用データ，テストデータに含まれる化学物質名全てを予測対象とするため，1 つのデータに集約した．さらに，データに α などの Unicode 文字が含まれているため，既存のツール[*1] を使用して ascii 文字へ変換した．

5.2 実験設定

4 節で学習した全てのモデルを使用して，5.1 節で準備した CHEMDNER のデータに対して予測を行った．その後，各クラスごととデータ全体に対して再現率を求め，それぞれのモデルでどのような性質が見られるかを比較した．

5.3 結果・考察

表 5 に結果を示す．太字の数字が各クラスに対して最も良い性能だった値を示し，下線の数字が最も性能が悪い値を示す．それぞれのデータセット内で評価した時と比較して，全体的に性能が大きく低下していた．その原因として，ChemIDplus に含まれる化学物質の種類は非常に多いものの，訓練に使用したのが正規化された見出し語のみであるため，記法のバリエーションが少なくなってしまったことが考えられる．そのため，見出し語だけでなく，同義語なども使用するなど工夫をする必要があると考えられる．

また，負例を整備した方が良い結果となっており，負例の選定が重要であることが確認された．例えば，図 1 の ethyl や methyl とそれらの下位語は，SYSTEMATIC の一部としてよく用いられるため，負例から取り除いたことで大きく性能が向上したと考えられる．その他にも，FAMILY や TRIVIAL，MULTIPLE などは，化学物質名に特徴的な文字列を持つものが多いため，負例から化学物質名を取り除いたことで再現率が向上したと考えられる．一方で，IDENTIFIER のようなデータベースで用いられる表現については，正例の割合が非常に大きいため，影響が小さかったと言える．

トークンの分割については，文字ベースのモデルよりも SentencePiece のモデルの方が良い結果となり，SentencePiece の中でも，語彙数が 4000 の方が良い結果となった．これは，SentencePiece を用いた場合，化学物質名に特有の接尾辞などをある程度まとまったトークンとして扱い，それらを化学物質名の特徴として学習することが期待されるためであると考えられる．語彙数による違いについては，語彙数が小さすぎる場合に特徴的なまとまりがあまり形成されないためであると考えられる．一方で，語彙数が大きすぎると，長いまとまりを 1 つのトークンとして扱ってしまうことで，数字などのわずかな違いをとらえられなくなることが懸念される．そのため，適切な語彙数を設定することは，識別システムの性能向上において重要であると考えられる．

[*1] https://github.com/spyysalo/unicode2ascii

表 5　異なる設定による訓練を行なった識別システムの CHEMDNER に対する再現率
Table 5　Recall of chemical named entities in CHEMDNER using different training settings of recognition system

	CID+WN			CID+WNchem			
クラス名	char	SP1k	SP4k	char	SP1k	SP4k	データ数
ABBREVIATION	0.409	0.435	0.373	0.389	**0.449**	0.443	1850
FAMILY	0.288	0.385	0.424	0.340	0.512	**0.571**	3544
FORMULA	0.791	0.825	0.840	0.836	0.848	**0.884**	1825
IDENTIFIER	0.981	**0.998**	0.991	0.993	0.997	0.997	577
MULTIPLE	0.502	0.562	0.635	0.643	0.637	**0.736**	518
NO CLASS	0.721	0.721	0.698	0.721	0.744	**0.767**	43
SYSTEMATIC	0.559	0.663	0.706	0.624	0.837	**0.885**	6638
TRIVIAL	0.312	0.344	0.390	0.341	0.467	**0.549**	4745
全体	0.469	0.537	0.567	0.514	0.654	**0.706**	19740

6. 論文中の表記を学習データに加えた実験

5.2 節では，ChemIDplus に登録されている単語が正規化されているため，記法のバリエーションが少ないという問題や，負例を整備することが重要であることがわかった．

この節では，CHEMDNER を訓練データに加えることで，記法のバリエーションが少ないという問題に対処することを試みる．

6.1 実験設定

5 節で使用した CHEMDNER のデータを，訓練用 8 割，開発用 1 割，テスト用 1 割となるように分割し，より性能の高かった負例を整備したデータセット追加して学習を行った．評価は，CHEMDNER のテスト用データを使用して再現率を比較した．

6.2 結果・考察

表 6 に結果を示す．CHEMDNER を訓練データに追加した結果，全てのモデルで性能が上がった．このことから，文中に用いられる表現を少量追加するだけでも，ChemIDplus のデータへの過学習をある程度抑えられることがわかった．また，SP1k が SP4k よりも再現率が高くなった他に，文字ベースのモデルの性能が大きく向上し，それぞれの性能差が小さくなった．しかし，4 節で行ったデータセットに対する評価（表 4）の際には，分割の違いによる差がほとんど見られなかったので，訓練時とデータの特徴が変わらなければ，どの分割法でも大きな差が見られないと考えられる．

SentencePiece の語彙数による違いについて，SP1k よりも SP4k の方が再現率が高くなった理由は，5.2 節で議論した通りである．これを踏まえると，訓練データに十分なバリエーションが確保できている場合は，比較的小さい語彙数で学習する方が良く，訓練データにあまり見られないバリエーションを扱うことが想定される場合は，比較的大きい語彙数で学習することが良いと考えられる．目的に応じて適切な大きさの語彙数を設定することが，識別システムの性能向上において重要であると言える．

7. おわりに

本稿では，化学物質名の多くが，命名規則によって生成されるという点に注目し，単語を文字列として分析することで化学物質か否かを識別するシステムを提案し，データベースに存在するような限られた事例については，非常に高い性能を持つことを確認した．一方で，実際の論文に出てくるような単語に適用したところ，あまり良い性能が得られず，表記の正規化が行われているようなデータベースからの正例の作成では不十分であることがわかった．また，負例についても，単純に，多くの語を集めれば良いわけではなく，事前に，化学物質名を含む可能性がある語を削除するといった処理が不可欠であることも確認した．また，サブワードへの分割の仕方がシステム全体の性能に影響を与えることも確認され，今後は，その影響について，より深く検討していく必要がある．

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] Huang, Z., Xu, W. and Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging, *CoRR*, Vol. abs/1508.01991 (online), available from ⟨http://arxiv.org/abs/1508.01991⟩ (2015).

[2] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016* (Knight, K., Nenkova, A. and Rambow, O., eds.), The Association for Computational Lin-

表 6 CHEMDNER を訓練に用いた識別システムの CHEMDNER に対する再現率
Table 6 Recall of chemical named entities in CHEMDNER using CHEMDNER for training of recognition system

	CID+WNchem			+CHEMDNER			
クラス名	char	SP1k	SP4k	char	SP1k	SP4k	データ数
ABBREVIATION	0.951	0.595	0.784	**0.957**	0.816	0.854	185
FAMILY	0.425	0.465	0.614	0.772	**0.842**	0.837	355
FORMULA	**0.973**	0.847	0.913	0.962	0.940	0.929	183
IDENTIFIER	1.000	1.000	1.000	1.000	1.000	1.000	58
MULTIPLE	0.635	0.654	0.712	0.923	**0.942**	**0.942**	52
NO CLASS	1.000	1.000	1.000	1.000	1.000	1.000	5
SYSTEMATIC	0.628	0.789	0.849	0.904	**0.947**	0.926	664
TRIVIAL	0.333	0.368	0.453	0.562	**0.709**	0.602	475
全体	0.595	0.620	0.713	0.812	**0.860**	0.829	1977

guistics, pp. 260–270 (online), DOI: 10.18653/v1/n16-1030 (2016).

[3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, (online), DOI: 10.1093/bioinformatics/btz682 (2019).

[5] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (online), DOI: 10.18653/v1/P16-1162 (2016).

[6] Akbik, A., Blythe, D. and Vollgraf, R.: Contextual String Embeddings for Sequence Labeling, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018* (Bender, E. M., Derczynski, L. and Isabelle, P., eds.), Association for Computational Linguistics, pp. 1638–1649 (online), available from ⟨https://www.aclweb.org/anthology/C18-1139/⟩ (2018).

[7] Leaman, R., Wei, C. and Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization, *J. Cheminformatics*, Vol. 7, No. S-1, p. S3 (online), DOI: 10.1186/1758-2946-7-S1-S3 (2015).

[8] Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H. and Wang, J.: An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition, *Bioinform.*, Vol. 34, No. 8, pp. 1381–1388 (online), DOI: 10.1093/bioinformatics/btx761 (2018).

[9] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).

[10] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada* (Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B. and Garnett, R., eds.), pp. 8024–8035 (online), available from ⟨http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library⟩ (2019).

[11] Gers, F. A., Schmidhuber, J. and Cummins, F. A.: Learning to Forget: Continual Prediction with LSTM, *Neural Computation*, Vol. 12, No. 10, pp. 2451–2471 (online), DOI: 10.1162/089976600300015015 (2000).

[12] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural BeygelzimerText Processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018* (Blanco, E. and Lu, W., eds.), Association for Computational Linguistics, pp. 66–71 (online), DOI: 10.18653/v1/d18-2012 (2018).

[13] National Library of Medicine: ChemIDplus, , available from ⟨https://chem.nlm.nih.gov/chemidplus/⟩ (accessed 2020-04-04).

[14] Miller, G. A.: WordNet: A Lexical Database for English, *Commun. ACM*, Vol. 38, No. 11, pp. 39–41 (online), DOI: 10.1145/219717.219748 (1995).

[15] Bird, S., Klein, E. and Loper, E.: *Natural Language Processing with Python*, O'Reilly (2009).

[16] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Bengio, Y. and LeCun, Y., eds.), (online), available from ⟨http://arxiv.org/abs/1412.6980⟩ (2015).

[17] Krallinger, M., Rabal, O., Leitner, F. et al.: The CHEMDNER corpus of chemicals and drugs and its annotation principles, *J. Cheminformatics*, Vol. 7, No. S-1, p. S2 (online), DOI: 10.1186/1758-2946-7-S1-S2 (2015).