Vol.2019-NL-241 No.7
2019/8/29

辞書を用いたコーパス拡張による化学ドメインのDistantly Supervised固有表現認識

辰巳 守祐[1,a)]　後藤 啓介[2,b)]　進藤 裕之[1,c)]　松本 裕治[1,2,d)]

概要：機械学習を用いた，科学領域などの専門分野の固有表現認識では，学習データのアノテーションコストが極めて高く，学習データの確保が課題である．この課題を解決するアプローチとして，Distant Supervision による自動アノテーションが注目されている．Distant Supervision を用いることで，辞書に登録されている化学物質名の出現箇所を Positive とする学習データをコーパスから機械的に作成できる．しかし，Distant Supervision に基づき作成されたデータにはノイズが多い．さらに，現実的に作成可能なサイズのコーパスが全ての化学物質名を網羅するのは一般的に困難であり，辞書に登録されているにも関わらず学習データに出現しない化学物質名が生じるという問題がある．そこで，本研究では Distant Supervision におけるノイズの影響の緩和と，辞書をより有効に利用する手法を提案する．人手アノテーションによる固有表現コーパスである ChemdNER で評価実験を行い，提案手法を用いて学習した固有表現認識モデルの有効性を確認した．

Data Augmented Distantly Supervised Named Entity Recognition in Chemical Domain

Shusuke Tatsumi[1,a)]　Keisuke Goto[2,b)]　Hiroyuki Shindo[1,c)]　Yuji Matsumoto[1,2,d)]

1. はじめに

近年，人手による精巧な素性選択を必要としない高性能な固有表現認識（Named Entity Recognition，以下 NER）モデルの研究が盛んに行われている [1–4]．しかし，これらのモデルの多くは学習に大量の人手アノテーションデータが必要である [2–5]．更に，化学や生命といった専門性の高いドメインでは，人手アノテーションのコストが高く，教師付きデータの確保が困難という問題がある．

アノテーションコストを削減するため，近年では Distant Supervision（以下 DS）による教師データの自動生成手法が注目されており，phrase mining [6], entity recognition [7–9], aspect term extraction [10], relation extraction [11] 等，様々なタスクで一定の成果を上げている．化学や生命などの専門分野では，専門用語の辞書が年々蓄積されており，大規模な辞書が無料で入手可能である．これらの辞書を利用してテキストコーパスに含まれる固有表現のマッチングを行うことで，人手のアノテーション無しに，NER の学習データを自動生成することできる．従来の DS NER では，単純に文字列マッチングによって教師データを生成する方法 [9,10] や，機械的な教師データ生成に伴うノイズを軽減させるために，品詞の正規表現に基づいて entity の範囲同定を行う手法 [7,8] などが提案されている．

化学ドメインにおける DS NER には，主に二つの問題がある．一つ目の問題は，化学辞書には一般用語が多く混在しているため，辞書マッチングによって機械的に生成された NER の教師データはノイズを多く含むという点である．表 1 に，化学物質で，かつ一般用語でもある単語の例を示す．例えば，‘water’ は化学ドメインにおいても一般用語として使用されることが多いが，化学物質名として使われる

1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
2 理化学研究所 革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project
a) tatsumi.shusuke.tk0@is.naist.jp
b) keisuke.goto@riken.jp
c) shindo@is.naist.jp
d) matsu@is.naist.jp

ⓒ 2019 Information Processing Society of Japan

表 1　化学物質名辞書 (高 Recall 辞書) と一般用語集 (Google 10,000 English) の積集合

‘nitrogen’, ‘timeline’, ‘universal’, ‘pain’, ‘butter’, ‘renaissance’, ‘perfume’, ‘diamond’, ‘disease’, ‘tea’, ‘tops’, ‘burns’, ‘obesity’, ‘oils’, ‘shock’, ‘fever’, ‘silicon’, ‘soil’, ‘oxygen’, ‘arthritis’, ‘elements’, ‘stroke’, ‘lead’, ‘tin’, ‘nickel’, ‘dust’, ‘amber’, ‘neon’, ‘glucose’, ‘eos’, ‘sodium’, ‘tramadol’, ‘ozone’, ‘chemicals’, ‘zinc’, ‘clay’, ‘calcium’, ‘hepatitis’, ‘coffee’, ‘coal’, ‘hydrogen’, ‘asthma’, ‘rid’, ‘brass’, ‘hydrocodone’, ‘titanium’, ‘diseases’, ‘minerals’, ‘quilt’, ‘mercury’, ‘cosmetics’, ‘carbon’, ‘acids’, ‘sas’, ‘gold’, ‘ice’, ‘cholesterol’, ‘insulin’, ‘phentermine’, ‘syndrome’, ‘smoke’, ‘confusion’, ‘infection’, ‘water’, ‘death’, ‘relate’, ‘lime’, ‘platinum’, ‘aluminum’, ‘austin’, ‘fortune’, ‘metals’, ‘silver’, ‘gambling’, ‘gasoline’, ‘petroleum’, ‘copper’, ‘apt’, ‘steel’, ‘ensure’, ‘latex’, ‘iron’, ‘steam’, ‘asbestos’

表 2　表記揺れサンプル

3-acetyloxy-4-(trimethylazaniumyl)butanoate
3-acetoxy-4-(trimethylammonio)butyrate
3-acetyloxy-4-(trimethylammonio)butanoate
3-acetoxy-4-(trimethylammonio)butanoate

こともあるため，化学辞書に登録されている．したがって，単純な文字列マッチングによって自動アノテーションを行うと，全ての ‘water’ が化学物質名としてアノテーションされてしまうという問題がある．そこで，本研究では，これらのノイズとなる固有表現を除去するために，NER の予測モデルに基づくノイズ除去手法を提案する (2.2 節)．具体的には，DS によって自動生成された教師データに対して，4-fold cross-validation による NER の予測を行い，自動アノテーションのラベルと予測ラベルが不一致の場合はノイズと見なし，このノイズを含むセンテンスを取り除く．

二つ目の問題は，略称や表記揺れにより学習データの Recall が低いという問題である．DS では，辞書に含まれない単語は基本的に NER の正解とならないため，そのような固有表現は False Negative となって学習のノイズとなる．一般に，化学ドメインでは略称や表記揺れが頻出するため，辞書のカバレッジを上げるには限界があり，False Negative を減らすことが難しい．例えば，‘3-acetyloxy-4-(trimethylazaniumyl)butanoate’ には，表 2 のような表記揺れが存在するが，辞書が全ての表記揺れを網羅するのは困難である．そこで本研究では，学習データの Recall を上げることを目的として，辞書を用いたコーパス拡張を提案する (2.3 節)．具体的には，DS によって生成されたデータ中の特定の単語と辞書単語を入れ替えることで，辞書単語を含む擬似センテンスを作成する．本稿では，このデータを Replaced データと呼ぶ．そして，この Replaced データを DS データに加えてコーパスの拡張を行う．これにより，NER モデルは DS データに登場しない化学物質名の文字系列も学習でき，未知語を含めた NER の性能向上が期待できる．

本研究では，化学ドメイン DS NER を辞書 Exact String Matching，‘BIOES’ 規格で行う．実験には，化学物質名辞書，論文アブストラクト，人手アノテーション評価データの 3 タイプのデータが必要である．化学物質名辞書は CTD (Comparative Toxicogenomics Database)[*1], MeSH (Medical Subject Headings)[*2], PubChem[*3]から収集する．論文アブストラクトは Medline（検索エンジンは PubMed[*4]）から収集する．Medline は生命医学系分野で最も使用されているオンラインデータベースで，文献毎に文献内容を示す，MeSH Term が 10~15 個程度付与されている．人手アノテーション評価データは ChemdNER [12] のテストデータを使用する．ChemdNER は PubMed で収集した論文アブストラクトに対して，化学分野の専門家が人手でアノテーションを行った，高品質な化学ドメイン NER コーパ

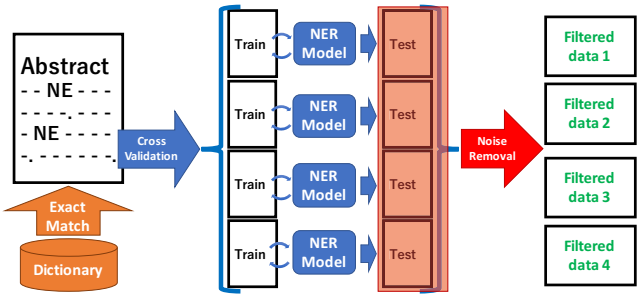

図 1　NER の予測モデルに基づく，Cross-Validation ノイズ除去．まず，DS データに対して 4-fold cross-validation を実行する．そして，DS に基づくアノテーションラベルを DS シルバーラベルと呼び，DS シルバーラベルと NER 予測ラベルが不一致の場合はノイズと見なし，このノイズを含む文を学習データから取り除く．

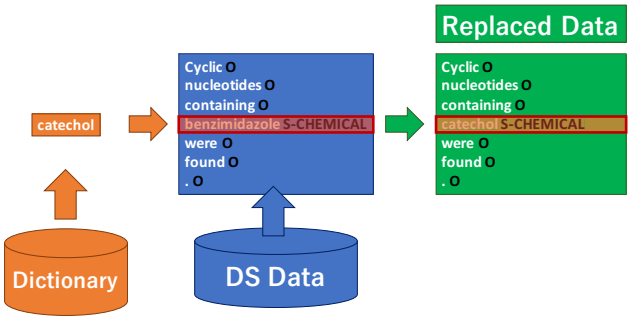

図 2　提案コーパス拡張に用いる Replaced データ作成プロセス．DS データ中の特定の単語と辞書単語を入れ替えた擬似的な文から構成されるデータを Replaced データと呼び，この Replaced データを DS データに加える．これにより，Replaced データ中には全ての辞書単語が登場するため，NER モデルは DS データには登場しない化学物質名の文字系列も学習できる．結果，未知語を含めた NER の性能向上が期待できる．

*1 http://ctdbase.org/
*2 https://www.ncbi.nlm.nih.gov/mesh
*3 https://pubchem.ncbi.nlm.nih.gov/
*4 https://www.ncbi.nlm.nih.gov/pubmed/

表 3 ‘acid’ を含む化学物質名サンプル

taurocholic **acid**
poly(N-isopropylacrylamide)-block-poly(aspartic **acid**)
2,4,6L-trinitrobenzene sulphonic **acid**
2,4,5-trimethoxycinnamic **acid**s

スである．主要化学分野を全て網羅しており，Train 3,500, Dev 3,500, Test 3,000 (アブストラクト数) から構成されている．この内，本研究では，Test データのみを使用する．

NER アーキテクチャは NER で主流の BiLSTM-CRF [13] を使用する．Word Embedding は化学ドメインの特徴である，固有表現定義の曖昧性より，全トークンを単語の文字情報だけで NER するのには限度がある．そこで，本研究では，文字情報に加えて，周辺文脈情報を Word Embedding に反映できる ELMo [14] を使用する．

また，化学物質名には特有のサブワードが存在する．例えば，‘acid’ という文字系列パターンは，化合物名の一部として頻出する (表 3)．この特徴を踏まえ，サブワード情報を考慮した Word Embedding である，FastText [15] も使用する．実験では，ELMo と FastText を Concatenate して用いる．

本研究の目標は，辞書を用いたコーパス拡張による，学習データの Recall 向上である．この目標達成の為，サブタスクを 3 つ定義する．

サブタスク 1 DS NER では，DS 学習データと人手評価データのドメインを揃えることが重要である．そこで，論文の出版年度とカテゴリを考慮した DS データの作成を行う．

サブタスク 2 DS データにはノイズが含まれるため，「NER の予測モデルに基づくノイズ除去手法」を提案する．

サブタスク 3 目的である学習データの Recall 向上のため，「化学物質名辞書を用いたコーパス拡張手法」を提案する．

最後に，本研究の主な貢献は次の 2 点である．

貢献 1 「NER の予測モデルに基づくノイズ除去手法」による，NER の F1 スコア向上．

貢献 2 「化学物質名辞書を用いたコーパス拡張手法」による，NER の F1 スコア向上．

2. 提案手法

2.1 DS データの作成

DS NER では，DS 学習データと人手評価データのドメインを揃えることが重要で，本研究では，論文アブストラクトの選定時，評価データである ChemdNER にドメインを揃える．まず，簡易な検索条件で収集した Large Data を作成する．次に，検索条件を更に絞り，ChemdNER により近い Selected Data を作成する．Medline 検索方法では，表 4 の通りに ChemdNER の各カテゴリに対応する MeSH Term を用いて検索する．Large Data は検索 MeSH Term でヒットした文献の内，1999.1.1~2018.12.31（20 年分）を使用する．一方，Selected Data はカテゴリのばらつき具合や文献の年度を ChemdNER に近づけるという目的の下，検索 MeSH Term でヒットした文献の内，2004.1.1~2013.12.31（10 年分）を使用する．更に，カテゴリ毎の上限を設け，年度順にソートし，年度の新しいものから上位（最大）1〜2 万文献を使用する．なお，対象を 2013 までにした理由は ChemdNER が 2013 年の文献を対象としているためである．選定した結果，表 5 のようになった．

化学ドメイン辞書には 2 つの特徴がある．1 つ目は固有表現の定義が曖昧なため，化学物質辞書に一般用語が混入しており，Precision が 100 にはならない点．2 つ目は，化学物質名の略称や表記揺れが多く，辞書カバレッジに限界があるため，Recall が低くなる点である．更に，Precision, Recall にはトレードオフの関係がある．そこで，比較検証の為，無料公開されている複数の辞書組み合わせを試し，評価データ（ChemdNER）に対する Precision が高かった組み合わせ（以下，組み合わせた辞書をマージして得られる単語の集合を高 Precision 辞書と呼ぶ）と Recall が高かった組み合わせ（以下，高 Recall 辞書）の 2 種類の辞書組み合わせを用意する．具体的には，化学分子データベースである，PubChem，MeSH(Medical Subject Headings)，CTD (Comparative Toxicogenomics Database) と一般用語集である Google 10,000 English[*5]の組み合わせの中で，最も Precision が高い辞書の組み合わせを高 Precision 辞書，最も Recall が高い辞書の組み合わせを高 Recall 辞書とする．

組み合わせを一通り試したところ，PubChem から一般用語集（Google 10,000 English）を取り除いた単語集合の Precision が最も高かった．一般用語を辞書から排除する動機は，化学物質名辞書には ‘water’ のように，化学物質名辞書に登録されているが，主として一般用語で使用される単語が混在しており，このような単語は Precision を下げる要因になるからである．実験では，文字数が顕著に長い化学物質名が実際の論文中に現れることは稀な為，PubChem 登録化学物質名の内，20 文字以下の化学物質名のみを使用する．以上の処理をした結果，辞書の規模は 128,252 単語，ChemdNER に対する Exact String Matching は Precision: 72.75, Recall: 12.39, F1: 21.18 となった．

PubChem，MeSH，CTD の組み合わせが最も Recall が高くなった．実験では，高 Precision 辞書の時と同様，20 文字以下の化学物質名のみを使用する．得られた辞書の規模は 238,758 単語，ChemdNER に対する Exact String Matching は Precision: 50.07, Recall: 53.66, F1: 51.80 となった．

*5 https://github.com/first20hours/google-10000-english

表 4　ChemdNER カテゴリと MeSH Term

ChemdNER のカテゴリ	Medline 検索条件 (MeSH Term)
PHARMACOLOGY & PHARMACY	Pharmacology
MEDICINAL CHEMISTRY	Chemistry, Pharmaceutical
ORGANIC CHEMISTRY	Chemistry, Organic
TOXICOLOGY	Toxicology
MULTIDISCIPLINARY CHEMISTRY	Interdisciplinary Studies
PHYSICAL CHEMISTRY	Chemistry, Physical
BIOCHEMISTRY & MOLECULAR BIOLOGY	Biochemistry
APPLIED CHEMISTRY	Chemistry, Clinical
ENDOCRINOLOGY & METABOLISM	Endocrinology
POLYMER SCIENCE	Polymers
CHEMICAL ENGINEERING	Chemical Engineering

表 5　各データのカテゴリ別文献数

Chem. Subject categories	Chemd	Selected	Large
PHARMACOLOGY	1,983	15,585	111,515
MEDICINAL CHEMISTRY	1,957	13,276	94,281
ORGANIC CHEMISTRY	1,893	1,366	2,237
TOXICOLOGY	1,664	4,059	8,557
MULTIDISCIPLINARY CHEMISTRY	1,217	299	713
PHYSICAL CHEMISTRY	997	11,886	51,391
BIOCHEMISTRY	879	12,064	320,109
APPLIED CHEMISTRY	843	701	1,273
ENDOCRINOLOGY	652	803	1,590
POLYMER SCIENCE	232	11,363	518,010
CHEMICAL ENGINEERING	3	167	342

DS データアノテーション方法については，辞書と論文アブストラクトから Exact String Matching アノテーションにより，自動作成する．作成手順は次の通り．

手順 1　辞書，および，論文アブストラクトをトークナイズする．単語のセグメンテーションについては「括弧，句読点，疑問符，感嘆符，引用符」と「先行，後続する単語」間にスペースを挿入する．ただし，括弧の前後のスペース情報を失わないようにするため，‘(’ の前，または ‘)’ の後が文字の場合はスペースを挿入しない．（表 6）．

手順 2　Exact String Matching アノテーションでは，論文アブストラクトの単語の内，辞書（化学物質名リスト）に含まれる単語に ‘BIES’ のいずれかを，その他の単語には ‘O’ をラベリングする．‘S’ は ‘Single’，‘B’ は ‘Begin’，‘I’ は ‘Inside’，‘E’ は ‘End’，‘O’ は ‘Outside’ を意味する．例えば，次の文において，‘hydroxy’ と ‘8-hydroxy decadienoic acid’ が固有表現の場合，ラベリングはこのようになる．‘Fluoro(O) and(O) hydroxyl(S) derivatives(O) of(O) 8-hydroxy(B) decadienoic(I) acid(E) were(O) prepared(O) .(O)’.

手順 3　化学物質名の含まれない文は学習データから除去する．辞書マッチアノテーションで得られるデータは負例（‘O’ ラベリングされた単語）の占める割合が高い為，NER モデルが正例（‘BIES’ ラベリングされた単語）を十分に学習できない．そこで，正例が含まれない文は学習データから除去し，正負例比率の偏りを軽減する．ただ，この処理により，DS データの文数が Exact String Matching アノテーションで使用する辞書の Recall によって増減する．具体的には，辞書の Recall が高くなるにつれて，正例数が増え，正例を含む文数も増えるため，学習データの文数も増加する．一方，辞書の Recall が低くなるにつれて，学習データの文数も減少する．

表 6　Word Segmentation のサンプル

例：‘(’ の前が文字であるため，スペースを挿入しない
処理前：Cu(II), Zn(II), and Pb(II) at pH 1.
処理後：Cu(II) , Zn(II) , and Pb(II) at pH 1 .

例：‘(’ の前，および，‘)’ の後が文字ではないため，スペースを挿入する
処理前：Recently, Investigational New Drug (IND) applications.
処理後：Recently , Investigational New Drug (IND) applications .

表 7　DS データサイズ（文数）

	高 Precision 辞書		高 Recall 辞書	
	Train	Dev	Train	Dev
Large	309,400	16,222	2,095,866	104,666
Selected	25,624	2,138	127,459	10,789

手順 4　ルールベースアノテーションを併用する．論文アブストラクト中の 20 文字以上の単語，及びハイフンを 3 つ以上含む単語は化学物質名である場合が多い為，‘BIES’ ラベリングする．以上の処理後，得られるデータサイズは表 7 のようになる．

2.2　NER の予測モデルに基づくノイズ除去

DS データには主に 2 種類のノイズがある．1 つ目は化学物質辞書に一般用語が混入することから生じる False Positive．2 つ目は辞書カバレッジの限界から生じる False Negative である．本研究では，これらのノイズを Cross-Validation により，NER の予測モデルに基づいて除去する．具体的には，DS データに対して 4-fold cross-validation を実行する（図 1）．そして，DS に基づくアノテーションラベルを DS シルバーラベルと呼び，DS シルバーラベルと NER 予測ラベルが異なる場合，次のような仮説を立てる．

仮説 1　NER の予測モデルの出力ラベルが False Positive（DS シルバーラベル：‘O’，NER 予測ラベル：‘BIES’）な場合，DS に基づくアノテーションが False Negative である可能性がある．

仮説 2　NER の予測モデルの出力ラベルが False Negative（DS シルバーラベル：‘BIES’，NER 予測ラベル：‘O’）な場合，DS に基づくアノテーションが False Positive である可能性がある．

これらの仮説に基づき，DS シルバーラベルと予測ラベルが不一致の場合はノイズと見なし，このノイズを含む文を取り除く．3.1 で作成した Selected Data に対してノイズ除去を行ったところ，処理後のデータサイズ（文数）は（表 8）のようになった．

2.3　辞書を用いたコーパス拡張

現実的に作成可能なサイズのコーパスが全ての化学物質

表 8　ノイズ除去，コーパス拡張後のデータサイズ（文数）

	高 Precision 辞書		高 Recall 辞書	
	Train	Dev	Train	Dev
Distant Supervision	25,624	2,138	127,459	10,789
+ Augmentation	153,876	2,138	366,459	10,789
Distant Supervision(ノイズ除去)	19,222	2,138	98,395	10,789
+ Augmentation(ノイズ除去)	147,474	2,138	337,395	10,789

表 9　Large Data と Selected Data の比較実験結果

	高 Precision 辞書			高 Recall 辞書		
	Pre	Rec	F1	Pre	Rec	F1
Large	62.79	18.91	29.07	55.77	64.37	59.76
Selected	**65.27**	**21.66**	**32.51**	**56.90**	**65.64**	**60.96**

名を網羅するのは一般的に困難であり，辞書に登録されているにも関わらず学習データに出現しない化学物質名が生じるという問題がある．この問題を解決するため，辞書を用いたコーパス拡張を提案（図 2）し，学習データの Recall 向上を試みる．具体的には，DS データ中の特定の単語と辞書単語を入れ替えた擬似的な文から構成されるデータを Replaced データと呼び，この Replaced データを DS データに加える．これにより，Replaced データ中には全ての辞書単語が登場するため，NER モデルは DS データには登場しない化学物質名の文字系列も学習できる．3.1 で作成した Selected Data に対してコーパス拡張を行ったところ，処理後のデータサイズ（文数）は表 8 のようになった．

3. 実験詳細

評価は人手評価データの ChemdNER を使用する．ChemdNER は 8 つの固有表現タイプから構成される．ただ，本研究では，FORMURA, ABBREVIATION, IDENTIFIER, MULTIPLE, NO CLASS は今回使用する辞書にほとんど含まれていないため，SYSTEMATIC, FAMILY, TRIVIAL のみを使用する（表 11）．具体的には，SYSTEMATIC, FAMILY, TRIVIAL のみを ‘BIES’ とラベルリングし，その他の FORMURA, ABBREVIATION, IDENTIFIER, MULTIPLE, NO CLASS は ‘O’ とラベリングする．

本実験では分散表現として ELMo と Fasttext を使用する．ELMo は化学ドメインのテキスト（Pubmed）で事前学習されたものが配布されているため，これを使用する．Fasttext は化学ドメインの配布がないため，自前で学習する．学習データは 3.1 の Large Data をトークナイズ, 小文字化し, 改行を除去したもの，1.5 億トークンを使用する．

3.1 DS データ作成の評価実験

3.1 の実験結果は表 9 のようになった．表 7 と表 9 より，Selected データの方が小規模ながら，Large データよりも NER の F1 が向上し，Data Distillation のような効果が見られた．出版年度とカテゴリの考慮は Distantly Supervised NER において重要だと分かった．

3.2 ノイズ除去，及びコーパス拡張の評価実験

3.2-3.3 の実験結果は表 10 のようになった．

観察 1　全体的に，Exact String Matching よりも Distant Supervision，及び提案手法の方が，Precision，Recall，F1 スコアが高かった．ただ，高 Precision 辞書の Precision については，Exact String Matching が最も高かった．

観察 2　NER の予測モデルに基づくノイズ除去が有効だと分かった．高 Precision 辞書，高 Recall 辞書いずれの実験においても，コーパス拡張の有無に関わらず，ノイズ除去により，F1 スコアが向上した．

観察 3　コーパス拡張については，高 Recall 辞書の実験において有効だった．高 Precision 辞書での実験では，ノイズ除去の有無に関わらず，コーパス拡張により F1 スコアが減少した．高 Recall 辞書での実験では，ノイズ除去の有無に関わらず，コーパス拡張により F1 スコアが向上した．

4. 考察

考察 1　高 Precision 辞書は Precision が高く DS データの False Negative ノイズが少ない．この結果，Replaced データ作成において，DS データ中の化学物質名 Entity と辞書単語を入れ替える際に，False Negative な化学物質名 Entity との入れ替えが減少し，精度の高い Replaced データが得られる．しかし，実験では，ノイズ除去有無に関わらず，コーパス拡張により全体的にスコアが悪化した．特に Recall の減少が顕著に見られた．高 Precision 辞書は約 13 万単語収録されているが，表 7 より，DS データが約 2.5 万文しかなく，Replaced データ作成時に同一の文が約 5 回，異なる辞書単語の入れ替え先として使用されている．（高 Recall 辞書の場合は，辞書が約 24 万単語，DS データが約 13 万文で，同一文の使用は 2 回程度）これにより，特定の化学物質名 Entity よりも，その周辺コンテキストの出現回数の方が多くなる．よって，NER モデルが Entity の文字情報よりも周辺コンテキストの特徴を学習し，推論時に Entity の文字情報よりもコンテキスト情報に過度に依存した予測を行うようになり，Recall が減少したものと思われる．

考察 2　高 Recall 辞書は Precision が低く，False Negative が多い．その結果，コーパス拡張により Precision が減少した．しかし，ノイズ除去併用時は，コーパス拡張により Precision が向上した．ノイズ除去により，False Negative な化学物質名トークンを除去できたものと思われる．

5. おわりに

本研究では，論文の出版年度とカテゴリを考慮した DS

表 10　ノイズ除去及びコーパス拡張の結果

	高 Precision 辞書			高 Recall 辞書		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Exact Match	**72.75**	12.39	21.18	50.07	53.66	51.80
Distant Supervision	65.27 ± 0.68	21.66 ± 1.44	32.51 ± 1.56	56.90 ± 0.38	65.64 ± 0.11	60.96 ± 0.26
+ Augmentation	64.59 ± 0.42	19.38 ± 0.18	29.81 ± 0.15	56.36 ± 0.26	66.60 ± 0.28	61.06 ± 0.25
Distant Supervision (ノイズ除去)	69.94 ± 0.90	**24.91 ± 0.32**	**36.74 ± 0.40**	56.71 ± 0.22	69.28 ± 0.48	62.37 ± 0.16
+ Augmentation (ノイズ除去)	71.77 ± 1.19	22.73 ± 0.30	34.52 ± 0.38	**57.23 ± 0.73**	**70.57 ± 0.36**	**63.20 ± 0.30**

表 11　ChemdNER のタイプ別 Entity 数と各辞書の Recall

タイプ	**CHEMDNER Test Entity 数**	**高 Precision 辞書 Exact Match Recall**	**高 Recall 辞書 Exact Match Recall**
SYSTEMATIC	**4,918**	**24.66**	**46.46**
FAMILY	**3,021**	**1.82**	**27.74**
TRIVIAL	**6,859**	**8.25**	**70.24**
FORMULA	3,234	0.00	0.00
ABBREVIATION	3,937	0.00	0.64
IDENTIFIER	475	0.00	10.53
MULTIPLE	165	0.00	0.00
NO CLASS	38	0.00	23.68

データである，Selected データの作成を行なった．実験 4.1 より，作成した Selected データが NER の性能向上に有効だと分かった．また，DS データにはノイズが含まれるため，NER の予測モデルに基づくノイズ除去手法を提案した．実験 4.2 の観察 2 より，当手法が有効だと分かった．最後に，辞書を用いたコーパス拡張を提案した．実験 4.2 の観察 3 より，高 Recall 辞書ではコーパス拡張が有効だと分かった．

今後の展望については，以下の 2 つが挙げられる．1 つ目は，辞書に対するノイズ除去．化学ドメインでは，辞書に一般用語が含まれている（表 1）ことから，そもそも辞書自体にノイズが混在しており，DS データ作成のボトルネックとなっている．2 つ目は化学物質名シソーラスを用いた，Replaced データの作成．現在は，Replaced データ作成時（図 2），単語の入れ替えをランダムで行なっているため，文脈のミスマッチが懸念される．そこで，化学物質名シソーラスを用いて，カテゴリの近い化学物質と入れ替えることで，ミスマッチの減少が期待できる．

謝辞

本研究は一部 JST CREST(課題番号: JPMJCR1513) の支援を受けて行った．富士通研究所の岩倉友哉氏には，PubChem の前処理済み辞書データを頂いた．また，Li Michael Wentao 氏には，Word Segmentation のスクリプトを，近藤修平氏，和田崇史氏，本多右京氏には，数々の助言を頂いた．ここに深く感謝の意を表する．

参考文献

[1] Li, J., Sun, A., Han, J. and Li, C.: A Survey on Deep Learning for Named Entity Recognition, *CoRR*, Vol. abs/1812.09449 (online), available from ⟨http://arxiv.org/abs/1812.09449⟩ (2018).

[2] Liu, L., Shang, J., Ren, X., Xu, F. F., Gui, H., Peng, J. and Han, J.: Empower Sequence Labeling with Task-Aware Neural Language Model, *ArXiv*, Vol. abs/1709.04109 (2017).

[3] Ma, X. and Hovy, E. H.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, *ArXiv*, Vol. abs/1603.01354 (2016).

[4] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *HLT-NAACL* (2016).

[5] Finkel, J. R., Grenager, T. and Manning, C. D.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *ACL* (2005).

[6] Shang, J., Liu, J., Jiang, M., Ren, X. D., Voss, C. R. and Han, J.: Automated Phrase Mining from Massive Text Corpora, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, pp. 1825–1837 (2017).

[7] Ren, X., El-Kishky, A., Wang, C., Tao, F., Voss, C. R. and Han, J.: ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering, *KDD : proceedings. International Conference on Knowledge Discovery and Data Mining*, Vol. 2015, pp. 995–1004 (2015).

[8] Fries, J. A., Wu, S., Ratner, A. and Ré, C.: SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data, *ArXiv*, Vol. abs/1704.06360 (2017).

[9] He, W.: Autoentity: automated entity detection from massive text corpora (2017).

[10] Giannakopoulos, A., Musat, C. C., Hossmann, A. and Baeriswyl, M.: Unsupervised Aspect Term Extraction with B-LSTM and CRF using Automatically Labelled Datasets, *WASSA@EMNLP* (2017).

[11] Mintz, M., Bills, S., Snow, R. and Jurafsky, D.: Distant supervision for relation extraction without labeled data, *ACL/IJCNLP* (2009).

[12] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., Sayle, R. A., Batista-Navarro, R. T., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K. H., Ramanan, S. V., Nathan, S., Žitnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S. A., Kors, J. A., Xu, S., An, X., Sikdar, U. K., Ekbal, A., Yoshioka, M., Dieb, T. M., Choi, M., Verspoor, K., Khabsa, M., Giles, C. L., Liu, H., Ravikumar, K. E., Lamurias, A., Couto, F. M., Dai, H.-J., Tsai, R. T.-H., Ata, C., Can, T., Usié, A., Alves, R., Segura-Bedmar, I., Martínez, P., Oyarzabal, J. and Valencia, A.: The CHEMDNER corpus of chemicals and drugs and its annotation principles, *Journal of cheminformatics*, Vol. 7, No. Suppl 1 Text mining for chemistry and the CHEMDNER track, pp. S2–S2 (online), DOI: 10.1186/1758-2946-7-S1-S2 (2015).

[13] 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).

[14] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M. and Zettlemoyer, L. S.: AllenNLP: A Deep Semantic Natural Language Processing Platform (2017).

[15] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146 (online), DOI: 10.1162/tacl_a_00051 (2017).

ⓒ 2019 Information Processing Society of Japan