修士論文

医学生物学論文からの情報抽出

渡邊 賢也

2019 年 3 月 8 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

渡邊 賢也

審査委員：

松本 裕治 教授　（主指導教員）
中村 哲 教授　（副指導教員）
新保 仁 准教授　（副指導教員）
進藤 裕之 助教授　（副指導教員）

医学生物学論文からの情報抽出*

渡邊 賢也

内容梗概

近年，医学生物学論文が日々多く投稿される中，機械学習を用いた情報抽出が期待されている．本研究では，BioCreative V Chemical Disease Relation dataset に対し，固有表現抽出，共参照解析，関係抽出を行い，論文中の情報を抽出する。医学生物学論文では，多くの関係が文をまたいでおり，関係抽出の際，文脈からより複雑な情報を抽出する必要がある．また，医学生物学論文に対するアノテーションには専門知識を要するので，学習データが小規模である事が多い．今回，固有表現抽出、関係抽出において精度の高く，文をまたぐ関係を抽出することができるモデルをベースにし，共参照解析モデルを追加し，各タスクの精度を向上させたモデルを提案する．医学生物学論文のタイトルと概要文に対し，固有表現を抽出し，全固有表現ペアに対し，共参照解析，関係抽出を行う．また，BioCreative V Chemical Disease Relation dataset 以外のデータを用いる事で，精度向上を試みた。

キーワード

固有表現，固有表現抽出，共参照解析，関係抽出，知識抽出，医学生物学，ニューラルネットワーク

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, 2019 年 3 月 8 日.

Information Extraction for Biomedical papers*

Kenya Watanabe

Abstract

In recent years, with the rapid increase in the volume of biomedical papers, information extraction using machine learning is expected. In this thesis, we perform named entity recognition, co-reference analysis, relation extraction on BioCreative V Chemical Disease Relation dataset and extract information in the abstract. In biomedical document, many relation across sentences. Thus, is necessary to infer relation from the context. In addition, annotating biomedical literature requires expert knowledge, and training data is often small scale. In this thesis, I use a base model that achieves high accuracy in named entity extraction, relationship extraction and can extraction relation across sentences. I propose enhancing the model with a module for co-reference analysis to further improve task performance. The resulting model extract named entities from the title and abstract of biomedical papers and perform co-reference analysis and relation extraction for all named entity pairs. Also, I improved performance by using external data.

Keywords:

Named Entity, Named Entity Recognition, Co-reference, Relation Extraction, Information Extraction, Biomedical Literature, Neural Network

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, March 8, 2019.

目次

図目次

第 1 章　序論

1.1　背景

近年．インターネットが普及し，日本語や英語といった自然言語で書かれた膨大な文書が誰でも入手可能になった．それらの文書の中から人間に有用な情報を取り出すことが重要視されている．

医学生物学においても，例えば，生命科学や医学生物学データベースである MEDLINE[52] では，約 40 万もの論文が毎年登録されている．この状況により，医学生物学の専門家や研究者でも医学生物学においての最新の研究成果を全て把握することが出来なくなっており，医学生物学論文から自動的に情報を抽出，構造化する試みが行われている．情報抽出には，固有表現抽出（Named Entity Recognition，NER）や共参照解析（Co-reference），関係抽出（Relation Extraction，RE）などがあるが，医学生物学論文に対してのアノテーションは専門知識を要し，学習データが小規模であることが多い．

固有表現抽出（Named Entity Recognition, NER）とは，固有表現（Named Entity, NE）を抽出し，クラスタリングする処理のことである．一般分野では，組織の名称・人名・地名などの固有名詞や，日時・温度・時間などの数値表現などを固有表現とする．また，専門分野では，化合物名や病名などの専門用語を固有表現とする．例えば，「Indomethacin induced hypotension in sodium and volume depleted rats.」という文に対して，「Indomethacin」，「hypotension」，「sodium」が固有表現であり，「Indomethacin」，「sodium」が「化学物質」に属し，「hypotension」が「病名」属すということを識別する処理を固有表現抽出という．固有表現は，次々と新しいものが出て来るため，全て辞書に登録し抽出するということが不可能であり，近年，機械学習を用いた固有表現抽出の手法が多く提案されている．

共参照解析（Co-reference）とは，ある文書が与えられたとき，文書内の具体的な実体を指す名詞句のうち，同じ実体を指す表現をグルーピングしてクラスタとしてまとめることである．例えば，「Scleroderma renal crisis (SRC) is a rare complication of systemic sclerosis (SSc) but can be severe enough to require temporary or permanent renal replacement therapy. 」という文において，「Scleroderma renal

crisis」,「SRC」,「systemic sclerosis」,「SSc」が固有表現であり,「病名」に属している．このとき，それぞれの固有表現が同じ固有表現かどうか，つまり,「Scleroderma renal crisis」,「SRC」が同じ固有表現であり,「systemic sclerosis」,「SSc」も同じ固有表現であることを識別する処理を共参照解析という．同じ固有表現であっても，上記のように略名が用いられることや，指示語が用いられることがある．略名や指示語が何を指しているかは文脈の情報を用いないとわからない場合がある．

関係抽出（Relation Extraction, RE）とは，文書集合から構造化された情報を抽出する情報抽出のタスクの一つであり，固有表現間の関係を表すような文脈から固有表現間の意味関係を抽出する処理である．例えば,「Indomethacin induced hypotension in sodium and volume depleted rats.」という文に対して,「Indomethacin」,「hypotension」,「sodium」が固有表現であり,「Indomethacin」,「sodium」が「化学物質」に属し,「hypotension」が「病名」属している．このとき,「Indomethacin」と「hypotension」が「原因・結果」の関係にあると識別する処理を関係抽出という．上記の例文のように，関係を抽出するには文脈の情報を用いなければならない．複数の文からなる文書に対して，文をまたいで関係を抽出する場合，より長く複雑な文脈の理解が必要とされ，難しい問題とされている．また，関係抽出をする場合，上記の例のように，固有表現が既知であるか，推定した上で，関係を推定することが多い．

固有表現抽出，共参照解析，関係抽出についての詳細な情報は二章で紹介する．

1.2 本論分の目的

本研究の目的は，医学生物学論文の概要文データである BioCreative V Chemical Disease Relation dataset（以後，CDR と記す）に対しての固有表現抽出，共参照解析，関係抽出において，精度向上を試みるとともに，より実問題に近い設定での実験を行う．本論文では，固有表現抽出，関係抽出のモデルである BRAN（Bi-affine Relation Attention Network）[32] をベースモデルとしたモデルを提案する．BRAN では，関係抽出を行う際に，固有表現抽出と共参照解析の正解データを用いているが，本研究では，より実問題の設定に近い，固有表現抽出と共参照解析の正解データがない場合にも関係抽出を行えるように BRAN を改良する．また，

外部データを用いて精度向上を試みる．

1.3　本論文の構成

本論文の構成は以下の通りである．2 章では，固有表現抽出や共参照解析，関係抽出の先行研究について紹介する．3 章では，ベースモデルと改良点について述べ，本研究の提案手法について説明する．4 章では，実験設定と実験結果について述べ，5 章で実験結果を考察する．6 章では，本論文の結論を述べる．

第 2 章　先行研究

本章では，固有表現抽出，共参照解析，関係抽出の先行研究を紹介する．

2.1　固有表現抽出

一般的に固有表現抽出は，文中の各単語に対し，定義した固有表現のクラスの中のどのクラスに属するか，一つの固有表現の範囲ラベルを付与する系列ラベリング問題として扱われる．ラベル付与方式として，BIO 方式（Begin，Inside，Outside），または BIOES 方式（Begin，Inside，Outside，End，Single）[1] が用いられる．以下の表に例文「... heparin induced thrombocytopenia type II ...」に対する系列ラベリングによる固有表現抽出の例を示す．

表 2.1　系列ラベリングによる固有表現抽出

文	...	heparin	induced	thrombocytopenia	type	II	...
BIO	...	B-Chemical	O	B-Disease	I-Disease	I-Disease	...
BIOES	...	S-Chemical	O	B-Disease	I-Disease	E-Disease	...

表 2.1 の例では，固有表現のクラスとして「Chemical」，「Disease」がある．単語ごとにその単語がどの固有表現のクラス（「Chemical」，「Disease」）に属し，固有表現のどの範囲（B-，I-）であるかを上記の様なラベルによって表現する．

上記の様な系列ラベリングによる固有表現抽出の手法には，大きく分けて，人手で作成した辞書などを用いるルールベースの手法と，機械学習を用いた手法の二つがある．

ルールベースを用いた手法は，人手で辞書などのパターンを作成し，それを適用することで固有表現抽出を行う [2]．この手法では，限られた分野においてルールに組み込まれている固有表現を高い精度で抽出することができる．また，人手でルールを作成するため，システムを利用する際，抽出規則が理解しやすい．しかし，背景で述べた様に，様々な分野において固有表現となる表現は日々増加しており，新たな固有表現に適応させるには，ルールを更新しなければならない．また，医学生

物学分野の様な分野の場合，専門的な知識を有していなければ，ルールの更新を行えず，ルール作成，更新に大きなコストがかかる．

機械学習を用いた手法は，人手でタグ付けされたテキストから学習し，固有表現抽出を行う．人手でタグ付けされたテキストを作成するためのコストはかかるが，一度学習したモデルを作成すると，未知の固有表現にも柔軟に適応されることが期待されている．機械機械学習を用いた手法には，最大エントロピー法（Maximum Entropy，ME）[3]，隠れマルコフモデル（Hidden Markov Model，HMM）[4]，条件付き確率場（Conditional Random Fields，CRF）[5]，サポートベクターマシーン（Support Vector Machine，SVM）[6]，ニューラルネットワークを用いた手法 [7][8][9] などが提案されている．ニューラルネットワークを用いた手法では，リカレントニューラルネットワーク（Reccurent Neural Network，RNN）をベースとし，条件付き確率場と組み合わせた手法 [10][11][12] が，CoNLL-2003 shared task database [†] などの既存の標準的な固有表現抽出のベンチマークデータセットに対して高い精度を示している．一般のテキストに対する固有表現抽出の研究は，Message Understanding Conference（MUC）[45] や Information Retrieval and Extraction Exercise（IREX）[46] などの会議で研究されてきた．

医学生物学分野のデータとしては，ChemdNER[51] や CDR[43] などがあり，MUC[45] 等で研究されてきた固有表現抽出の技術を適用する研究が行われてきた．また，表 2.1 の様に系列ラベリングによる固有表現抽出を行う場合は，文を単語ごとの様にトークンに分割してそれぞれのトークンについて識別を行うが，医学生物学分野などの分野において固有表現抽出を行う場合，文を単語ごとではなく部分単語（subword）に分割することがある [53]．これは，化学物質名や病名などの固有表現は似た綴りを含むことや部分的な表現を組み合わせていることが多いためである．例えば，「headache」や「stomachache」には同じ「-ache」という綴りが含まれており，「1,2-dimethylhydrazine」のような「,」や「-」などで部分的な表現を組み合わせている表現が多く存在する．この様な部分的な綴りや表現の情報を効果的に用い，規則性のある未知語に適応するため，単語を部分単語に分割して用いる．また，本来一般分野のデータに対しては，スペース区切りの単語を一つのトークンとして扱い，系列ラベリングによる固有表現抽出を行う．しかし，ChemdNER[51]

[†]https://www.clips.uantwerpen.be/conll2003/ner/

などのデータでは，固有表現がスペースでは区切れない場合が少数存在する．医学生物学分野のデータに対しての単語分割についての研究も行われている．単語から部分単語への分割方法としては，統計量を用いた分割方法が機械翻訳の研究にも多く用いられている．本研究で用いる byte pair encoding(BPE)[13] は，元々データ圧縮の分野で提案された手法であり，テキストの圧縮率を目的関数にして，どん欲的に分割を決定する部分単語分割アルゴリズムである．

また，本研究で用いる CDR に対しての固有表現抽出としては，tmChem[35] や DNorm[36][37]，条件付き確率場を用いた手法 [39] などがある．

2.2 共参照解析

共参照解析では，CoNLL 2012 shared task[47] など一般分野のデータやこのデータに対するモデルには多いが，医学生物学分野のような専門分野での大規模データや分野に特化したモデルは多くない．評価としては，近年，MUC[45]，B^3[49]，$CEAF_e$[50] の三つとそれらの平均である CoNLL score を指標とするのが主流である．MUC[45] は，予測したメンションによるクラスタと，実際のコーパスのクラスタの境界を使って部分集合を作り，共参照解析のスコアを算出する．B^3[49] は，正解クラスタと予測のクラスタのメンションの重なりの度合いで共参照解析のスコアを計算する．$CEAF_e$[50] は，クラスタの固有表現同士でのアライメントを取り，同じ固有表現を示すクラスタ同士でメンションに基づく類似性により算出する．CoNLL score は，以上のスコアを用いて計算されている [50]．本研究での，共参照解析の評価は 4 章で説明する．

共参照解析の手法にも，大きく分けて，人手で作成した辞書などを用いるルールベースの手法と機械学習を用いた手法の二つがある．

ルールベースの共参照解析モデルとしては，複数の Sieve と呼ばれるルールを精度の高い順に適用しながら共参照解析をを行うモデル [14] などがある．このモデルは，CoNLL 2011 shared task の Closed Track で最も高い精度を出し，CoNLL score で 57.79 となっている．

機械学習を用いた共参照解析モデルには，大きく分けて，Mention Pair モデル [15][16][17] や Mention Ranking モデル [18][19]，または複数のモデルを組み合わ

せたもの [54] などがある．

Mention pair モデルでは，最初にルールなどの抽出されたメンションを用いて，その中の任意の二つのメンションに対して共参照の関係があるかどうかを個別に判定する．それぞれの照応詞で先行詞として最も適しているものを選んでクラスタリングすることが多い．しかし，共参照のクラスタを作る過程で一貫性が保てないという問題点がある．また，先行詞と照応詞の間にどれくらい関係があるかを考慮できるが，ある照応詞に対してどの先行詞が最も共参照の関係にあるかという可能性に関して考慮できない．この問題を解決するためのモデルが Mention Ranking モデルである．

Mention Ranking モデルは，文書を頭から順方向に処理しながら，それぞれのメンションについて先行詞をランキングすることによって先行詞を決定するモデルである．局所的なペアだけでなく，文書全体のメンション の先行詞の候補を考慮に入れることができる．しかし，先行詞の中でもっとも共参照関係にあると思われるものを選べても，クラスタの一貫性を保つことはできない場合が多いと指摘されている [54]．

先述のモデルを組み合わせた手法を提案している論文 [54] では，クラスタリング時の誤りを減らすため，メンションをクラスタリングする過程で，それまでに作られたクラスタの情報を用いた上で，B^3 と MUC スコアに対して最適化するよう DAgger を使って学習することで，クラスタ内の実体の一貫性を保つ手法がを提案している．また，ニューラルネットを用いた手法としては，Mention Ranking モデルを学習するモデル [21] やクラスタの一貫性を保つためにリカレントニューラルネッ トワークを用いているモデル [22] などがある．

また，本研究で用いる CDR に対しての共参照解析としては，外部データを用いた辞書マッチングを用いた手法 [40][41]，ベクトル空間モデルを用いた手法 [39]，tmChem[35] と DNorm[36][37] などがある．

2.3 関係抽出

本論文では，二つの固有表現間の関係を抽出する関係抽出タスクについて述べる．以下で関係抽出タスクの手法として，教師あり学習や教師なし学習，半教師あり学

習，Distant Supervision を用いたモデルを紹介する．

教師あり学習のデータとして The NIST Automatic Content Extraction（ACE）RDC 2003 and 2004 コーパス [48] などがある．教師あり学習の手法として，サポートベクターマシンを用いた手法 [23]，パーセプトロンを用いた手法 [24]，文脈依存情報を用いた木構造による木カーネルベースを用いた手法 [25] などの研究が行われている．教師あり学習は，安定して高い精度を出すことが多いが，学習データを作成するコストが高く，データ量が制限されるという問題がある．

教師なし学習では，大量のテキストデータ，関係を扱うことが出来る．例えば，大量のテキストから固有表現間の単語の文字列を抽出，処理し，新たな関係ラベルが付与されたデータを抽出する研究が行われている．またシステムとして，関係の集合を抽出することが出来るオープン関係抽出システム [26] や，関係のある文書から教師なし関係抽出とクラスタリングを用いることで，情報抽出をする IDX システム [27] などがある．教師なし学習を用いた関係抽出には，教師なし学習によって得られた関係ラベルと，人手で作成された関係ラベルの対応づけが難しいなどの問題がある．人手で作成された関係ラベルとは，例えば知識ベースにあらかじめ定義されているラベルなどである．

半教師あり学習を用いた手法として，ブートストラッピング法を用いた手法がある．関係抽出におけるブートストラッピング法 [28] は，以下の手順を繰り返す事で少数のインスタンスから大規模なインスタンスを再帰的に獲得する手法である．

- 獲得対象となる関係クラスのインスタンスをシードとして与える
- テキストコーパスからインスタンスと共起するパターンを抽出する
- 抽出した共起パターンを用いて新たなインスタンスを抽出する

しかし，ブートストラップ法には，反復しているとシードのインスタンスとは全く関係のないインスタンスを抽出してしまうという問題がある．

Distant Supervision は，上記の手法の利点を組み合わせた手法 [29] であり，医学生物学分野の弱教師あり学習を用いた手法 [30][31] に近い設定である．Distant Supervision は，Freebase[44] などの知識ベース呼ばれる大規模データベースと NYTimes などのコーパスを用いて擬似的な正解ラベルを作成する．固有表現を二つ以上含む文をテキストコーパスから収集し，知識ベースに含まれる固有表現間の

関係ラベルを擬似的な正解ラベルとして付与する事で学習データを作成する．擬似的な正解ラベルのついたデータを作成することで，教師あり学習よりも学習データが多くなる．これは，教師あり学習のデータ不足の問題の解決法の一つとなる．教師なし学習の手法との相違点は，知識ベースに含まれる固有表現間の関係ラベルを擬似的な正解ラベルとして付与する点である．擬似的な正解ラベルを与えることで，「教師なし学習によって得られた関係ラベルと人手で作成された関係ラベルの対応づけ」の問題を解決することができる．ただし，実際にはラベルが付与されるべきではない文に対してもラベルを付与することがあるが，大量のデータを獲得し，複数のインスタンスを考慮しラベル付けを行えば，そういった場合を無視して学習することができると言われている．

また，本研究で用いる CDR に対しての関係抽出としては，構文木などの特徴量をとりニューラルネットを用いた手法 [38]，サポートベクターマシーンを用いた手法 [39][40]，ルールベースの手法 [41] などがある．

[38] では，特徴量ベースモデルとツリーカーネルモデル，ニューラルネットワークモデルの識別スコアの重み付き和により関係抽出を行う．この手法では，単文で「Chemical」と「Disease」の両方の固有表現を含む文に対して関係抽出を行う．特徴量ベースモデルは，特徴量として文脈，固有表現，位置，距離，動詞を特徴量とし，多項式カーネルを用いて識別スコアを計算する．単語，語幹，品詞，二つの固有表現の前後三単語のチャンクを文脈の特徴量，ヘッド，品詞，チャンクの情報を固有表現の特徴量，「Chemical」の固有表現と「Disease」の固有表現のどちらが前にあるかという情報と位置の特徴量，二つの固有表現間に幾つの単語があるかという情報を距離の特徴量，動詞が二つの固有表現の前，間または後のいづれにあるかという情報を動詞の特徴量として用いる．ツリーカーネルモデルは，単語，品詞，構文木の情報を特徴量とし，カーネル法を用いて識別スコアを計算する．ニューラルネットワークモデルは，単語列と品詞列，固有表現をヘッドワードに置き換えた文，構文木の情報を LSTM（Long short-term memory）を用いて処理し，識別スコアを計算する．

[39] では，文脈や固有表現，事前知識の情報を特徴量とし，サポートベクターマシーンを用いて関係抽出を行う．文脈や固有表現，事前知識の情報は以下のような情報がある．

- 文脈情報
  - 固有表現の前の Bag-of-words と bigram
  - 固有表現の間の Bag-of-words と bigram
  - 固有表現の後の Bag-of-words と bigram
  - 二つの固有表現が同じ文に含まれるかどうか
  - 二つの固有表現が隣接した文に含まれるかどうか
  - 二つの固有表現が二文以上離れた文に含まれるかどうか
  - 固有表現間の単語に特有のキーワードが含まれているかどうか
- 固有表現情報
  - 固有表現の Bag-of-words と bigram
  - 「Chemical」と「Disease」のどちらが先に出現するか
  - 固有表現の MeSH IDs
  - 「Chemical」の固有表現がタイトルに含まれているまたは概要文に最も多く含まれる主要な固有表現かどうか
- 事前知識の情報
  - MeSH（Medical Subject Headings）[57] における「Chemical」,「Disease」の固有表現のカテゴリー
  - MeSH における特定の「Disease」の固有表現が文に含まれているか
  - MeSH における一般的な「Disease」の固有表現が文に含まれているか
  - 「Chemical」,「Disease」の固有表現ペアの MEDI（MEDication Indication Resource）[55] の情報
  - 「Chemical」,「Disease」の固有表現ペアの SIDER（Side Effect Resource）[56] の情報
  - 「Chemical」,「Disease」の固有表現ペアの CTD[58] の情報

[40] では，言語情報や統計情報，事前知識の情報を特徴量とし，サポートベクターマシーンを用いて関係抽出を行う．言語情報は，固有表現ペアが文章中で同じ文に含まれてるかどうかなどの情報と構文木の情報である．統計情報は，全ての固有表現ペアが文章中で出現した回数や固有表現間の文や単語の数，タイトルに出現したかどうか，固有表現が MeSH に含まれているかどうかの情報である．事前知識の情報は，UniProt や CTD，UMLS などの構造化されたデータベースや科学論

文の概要文から Euretis Knowledge Platform を用いて得た情報である．

[41] では，「Chemical」と「Disease」の固有表現の前後の特定の単語とフレーズのパターンを設定し関係抽出を行なう．例えば，「Chemical」の固有表現の後に「caused」という単語があるかどうかなどのパターンを設定する．

第 3 章　提案手法

本章では，本論文で提案するモデルのベースモデルとベースモデルからの改良点について説明する．

3.1　ベースモデル

本研究で提案するモデルのベースモデルである BRAN[32] は，固有表現抽出と関係抽出を行うモデルである．BRAN は，論文の概要文の様な長い文章に対して文脈情報を用いて関係抽出を行う構造となっている．また，2 章で紹介した部分単語を用いることで，医学生物学論文に含まれる医学生物学用語を抽出しやすくしている．BRAN を提案した論文では，部分単語の作成として，BPE[13] が用いられている．

BPE は，テキストの圧縮率を目的関数にして，貪欲的に分割を決定していく部分単語分割アルゴリズムである．まず文字単位で部分単語とし，共起しやすいトークンを結合していくことで，単語を出現率の高い部分単語に分割する．

次に，BRAN の詳細な説明に移る．トークンの語彙数を m，トークンの系列のとりうる長さを l，入力されたトークンの系列の長さを N とする．トークンの埋め込みベクトルと位置ベクトルは同じ次元であり d とする．トークン埋め込み行列は $O^{m\times d}$，位置埋め込み行列は $P^{l\times d}$ とする．モデルの入力を x とすると，i 番目のトークンの入力 x_i は，以下の式の様に，i 番目のトークンの埋め込みベクトル s_i と i 番目の位置の埋め込みベクトル p_i の和で表される．

$$x_i = s_i + p_i \tag{3.1.1}$$

トークンのエンコーダには，Transformer self attention model（Transformer）[33] をベースにしたモデル（以後，TF と記す）を用いる．TF は，同じ構造のモデルを B 層積み重ねて構成されている．ただし，それぞれの層で異なるパラメータを持つ．TF の出力を b_i とすると，k 層目の TF_k の出力 $b_i^{(k)}$ は，$k-1$ 層目の出力 $b_i^{(k-1)}$ と $k-1$ 層目の出力 $b_i^{(k-1)}$ を TF_k に入力した結果の和で，以下の様な式で

表される．ただし，$b_i^{(0)} = x_i$ とする．

$$b_i^{(k)} = b_i^{(k-1)} + \mathrm{TF}_k(b_i^{(k-1)}) \tag{3.1.2}$$

TF には，Multi-head attention と畳み込み層（Convolutional Neural Network，CNN）がある．

Multi-haed attention は，個別の正規化されたパラメータを用いて，self-attention を同じ入力に対して複数回適用する．キー k，値 v，クエリ q，とする．k，v，q はそれぞれ d/H 次元である．ここで H はヘッドの数である．ヘッド h の a_{ijh} は以下の式で表される．

$$a_{ijh} = \sigma\left(\frac{q_{ih}^T k_{jh}}{\sqrt{d}}\right) \tag{3.1.3}$$

σ は j のディメンションでのソフトマックスを意味する．個々のアテンションの出力は以下の o で表す．

$$o_{ih} = \sum_j v_{hj} \bigodot a_{ijh} \tag{3.1.4}$$

$$o_i = [o_{i1}; ...; o_{ih}] \tag{3.1.5}$$

$\bigodot$ は要素ごとの積を意味し，$[\cdot;\cdot]$ は結合を意味する．Multi-head attention の入力と出力の和を Layer normalization（LN）[34] したものを z とする．

$$z_i = \mathrm{LN}(b_i^{k-1} + o_i) \tag{3.1.6}$$

Layer normalization は，入力 *input*，出力 *output* に対して以下の式のように計算する．

$$input_{mean} = \frac{1}{d}\sum_{i=1}^{d} input_i \tag{3.1.7}$$

$$input_{var} = \left(\frac{1}{d}\sum_{i=1}^{d} input_i \bigodot input_i\right) - \left(input_{mean} \bigodot input_{mean}\right) \quad (3.1.8)$$

$$norm = \frac{input - input_{mean}}{\sqrt{input_{var} + \epsilon_{LN}}} \quad (3.1.9)$$

$$output = LN_{parameter_g} \bigodot norm + LN_{parameter_b} \quad (3.1.10)$$

ただし $LN_{parameter_g}$ と $LN_{parameter_b}$ は d 次元で定義し，式 3.1.10 では $norm$ に次元が合うように要素を複製，拡張する．

続いて畳み込み層について説明する．Transformar を提案した論文では，ウィンドウ幅 1 の畳み込み層を二つ積み重ねていたが，TF では，それに加えウィンドウ幅 5 の畳み込み層を追加する．これは，ローカルな文脈情報を取るための構造である．関係抽出を行う際に，このようなローカルな文脈情報が有用であると考えられる．ウィンドウ幅 w の畳み込み層を $C_w(\cdot)$ と表すと，畳み込み層の計算は以下のようになる．

$$t_i^{(0)} = ReLU(C_1^{(0)}(z_i)) \quad (3.1.11)$$

$$t_i^{(1)} = ReLU(C_5(t_i^{(0)})) \quad (3.1.12)$$

$$t_i^{(2)} = C_1^{(1)}(t_i^{(1)}) \quad (3.1.13)$$

ここまでのモデルは，固有表現抽出と関係抽出，両タスクにおいて同じである．

固有表現抽出は，TF の出力 $b_i^{(B)}$ を固有表現クラス n の次元に写像する．$d \times n$ 次元の $W^{(0)}$ を用いて，固有表現のそれぞれのクラスのスコアを $score_{ner}$ とすると，以下のような式になる．

$$score_{ner} = W^{(0)} b_i^{(B)} \quad (3.1.14)$$

このスコアの最も高い固有表現クラスを予測ラベルとする．訓練時の損失関数はクロスエントロピーを用いる．損失関数は以下のような式で計算される．

$$\frac{1}{N}\sum_{i=1}^{N}\log P(n_i \mid score_{ner}) \tag{3.1.15}$$

BRAN で行う関係抽出における関係には，CDR で抽出したい関係のように＜head，tail ＞があると想定している．CDR で抽出する関係は，Chemical-induce Disease であり，病気と病気を誘発させる化学物質間の関係を抽出する．このとき，head を化学物質，tail を病気とする．TF の出力 $b_i^{(B)}$ を関係の head と tail それぞれについて以下の式のようにエンコードする．

$$e_i^{head} = W_{head}^{(2)}(\mathrm{ReLU}(W_{head}^{(1)}b_i^{(B)})) \tag{3.1.16}$$

$$e_i^{tail} = W_{tail}^{(2)}(\mathrm{ReLU}(W_{tail}^{(1)}b_i^{(B)})) \tag{3.1.17}$$

式（3.1.12，3.1.13）の W は全て同じ次元で，$d \times d$ である．

次に，bi-linear を用いて，全てのトークンペアに対しての関係ラベルに対するスコアを計算する．関係ラベルが L 種であり，$N \times N \times L$ 次元のテンソル A を用いて以下の式のように計算される．

$$A_{ilj} = (e_i^{head}L)e_j^{tail} \tag{3.1.18}$$

次に，ある head の固有表現のトークンと tail の固有表現のトークンのペアに対しての関係ラベルに対するスコアを LogSumExp 関数を用いて集計することで，ある head の固有表現と tail の固有表現のペアに対しての関係ラベルに対するスコアを以下の式のように計算する．

$$scores(p^{head}, p^{tail}) = \log \sum_{\substack{i \in P^{head} \\ j \in P^{tail}}} \exp(A_{ij}) \tag{3.1.19}$$

この操作は，固有表現抽出と共参照解析の正解データまたは予測データを用いて行う．BRAN を提案した論文では，は関係抽出を行う際に，固有表現抽出と共参照

解析の正解データを与えることを想定している．また，このスコアを元に閾値を用いて予測ラベルを生成する．

訓練時の損失関数はクロスエントロピーを用いる．損失関数は E 個の固有表現ペアでは，以下のような式で計算される．

$$\frac{1}{E}\sum_{i=1}^{E}\log P(r_i \mid scores(p^{head}, p^{tail})) \tag{3.1.20}$$

最適化アルゴリズムは，Adam を用いており，学習率は 0.0005 で，clip grafients to norm 10 とする．また，5 割の確率で固有表現抽出を，5 割の確率で関係抽出を学習する．また，関係抽出を学習する際には，negtive data と positive data をそれぞれ 5 割の確率で学習する $B = 2$ で，$H = 4$ で，バッチサイズは 32 である．また，token の 15% を無作為に UNK トークンに置き換える．また，本論文で用いる CDR に対して行うとき，gradient noise 0.1 で与え，トークンの埋め込み行列を PubMed に含まれる概要文の中から 10% の概要文を用いて学習する．トークンの埋め込み行列の学習には，skipgram を用いる．また，Adam のハイパーパラメータ ϵ を 0.0001，β_1 を 0.1，β_2 を 0.9，token のドロップアウト率を 0.15，o のドロップアウト率を 0.15，A のドロップアウト率を 0.65，ϵ_{LN} を $1e^{-8}$ に，$LN_{parameter_g}$ を要素 1 の d 次元のベクトル，$LN_{parameter_b}$ を要素 0 の d 次元のベクトルと設定する．

BRAN 全体の処理の図 3.1 に示す．

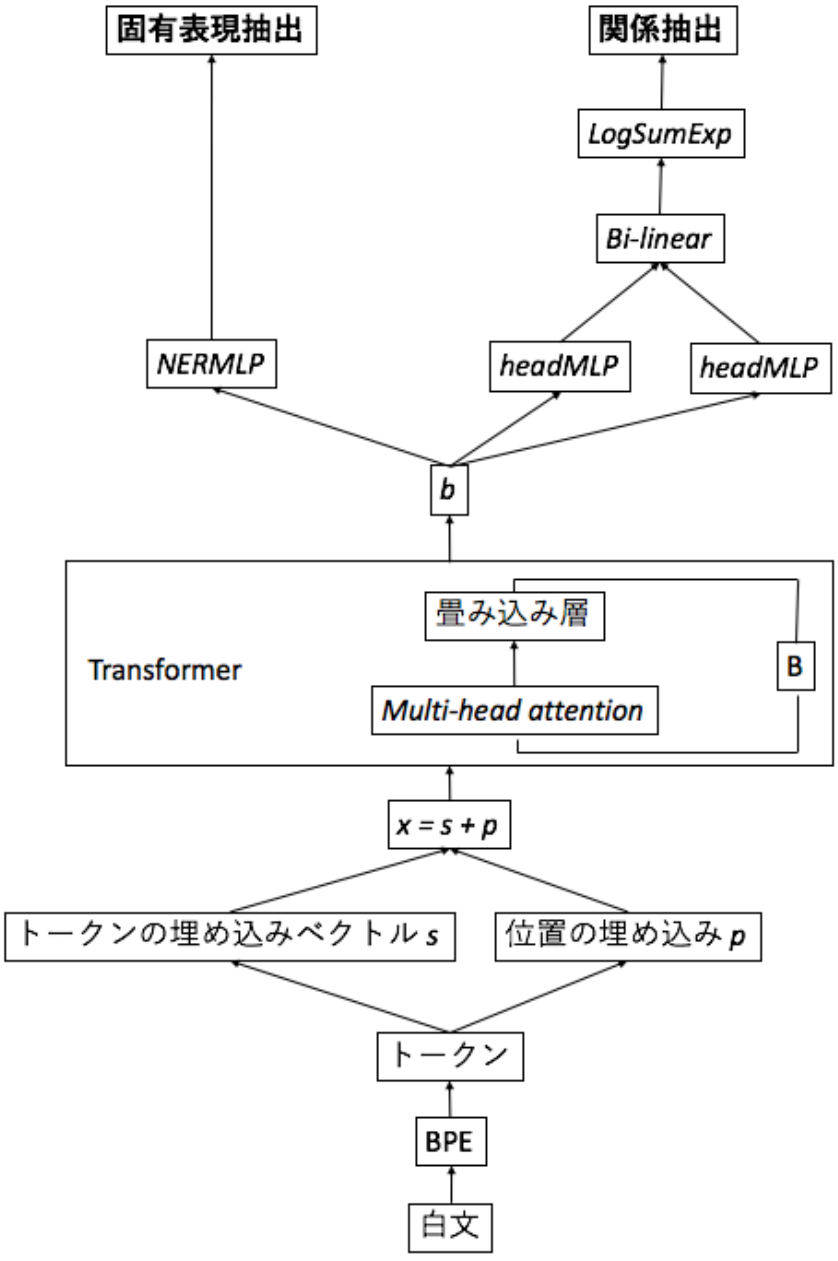
固有表現抽出
関係抽出
LogSumExp
Bi-linear
NERMLP
headMLP
headMLP
b
Transformer
畳み込み層
Multi-head attention
B
x = s + p
トークンの埋め込みベクトル s
位置の埋め込み p
トークン
BPE
白文

図 3.1　BRAN

3.2 改良点

提案手法は，3.1 節で紹介した BRAN を改良したモデルを用いる．以下で，改良点について説明する．改良点としては，共参照解析モデルの追加，固有表現抽出モデルの変更がある．共参照解析のタスクを増やしたことによる訓練時の設定も変更は 4 章で述べる．

共参照解析モデルは，本研究と同じ分野のデータに対して研究され，BRAN に組み込みやすいニューラルネットのモデル [54] を参考に作成した．この共参照解析モデルは，ある概要文中に含まれる固有表現の全ての組み合わせにおいて，固有表現ペアを作成し，ペアに含まれる固有表現が互いに同じ固有表現であるかどうかを推定する．入力は，ある文章に含まれる任意の固有表現ペアそれぞれのトークンの埋め込みベクトルで，出力は入力された固有表現のペアが互いに同じ固有表現であるかどうかの二値のスコアである．

固有表現ペア（$entity_1$，$entity_2$）のトークンの埋め込みベクトルをそれぞれ s_{entity_1}，s_{entity_2} と表す．例えば，$entity_1$ が文章の i 番目から j 番目のトークンであるとき，$s_{entity_1} = [s_i; ...; s_j]$ であり，$(j - i) \times d$ 次元となる．

s_{entity_1}，s_{entity_2} を入力とし，それぞれに対して畳み込み層，Maxpooling 層を用いて，以下の式のようにエンコードする．ただし，3.1.11 式と同様にウィンドウ幅 w の畳み込み層を $C_w(\cdot)$ と表し，Maxpooling 層を $Maxpool(\cdot)$ と表す．また，畳み込み層，Maxpooling 層の出力を $cnns$ とする．

$$cnns_{entity_1} = Maxpool(C_5^{(entity_1)}(s_{entity_1})) \tag{3.2.1}$$

$$cnns_{entity_2} = Maxpool(C_5^{(entity_2)}(s_{entity_2})) \tag{3.2.2}$$

次に，Maxpooling 層の出力を結合する．また，このとき cnn_{entity_1} と cnn_{entity_2} のコサイン類似度も算出し結合する．コサイン類似度の操作を，$cossim(\cdot, \cdot)$ と表す．

$$cnns_{concat} = [cnns_{entity_1}; cnns_{entity_2}; cossim(cnns_{entity_1}, cnns_{entity_2})] \tag{3.2.3}$$

コサイン類似度の操作は，入力ベクトルを $input_1$，$input_2$，出力値を $output$ とすると以下のような式で計算される．

$$output = \frac{input_1 \cdot {input_2}^T}{\sqrt{\sum_{i=1}^{d} input_1 \bigodot input_1 \times \sum_{i=1}^{d} input_2 \bigodot input_2}} \tag{3.2.4}$$

次に，以下のように固有表現ペアが互いに同じ固有表現であるかどうかの二値のスコアを計算する．ただし，$W^{(3)}$ は $(2d+1) \times (2d+1)$ 次元，$W^{(4)}$ は $(2d+1) \times (2)$ 次元である．

$$scores_{coreference} = W^{(4)}(W^{(3)} cnns_{concat}) \tag{3.2.5}$$

二値のうち大きいスコアを予測ラベルとする．訓練時の損失関数はクロスエントロピーを用いる．損失関数は E 個の固有表現ペアでは，以下のような式で計算される．

$$\frac{1}{E} \sum_{i=1}^{E} \log P(y_i \mid score_{coreference}) \tag{3.2.6}$$

続いて，固有表現抽出モデルの改良点について説明する．改良点は，BRAN の固有表現抽出モデルの出力 $score_{ner}$ に対し，条件付き確率場を適応する点である．ただし，CRF の操作を $CRF(\cdot)$ と表わす．

$$score_{ner}^{new} = CRF(score_{ner}) \tag{3.2.7}$$

モデルにおけるその他の操作は，同じである．ただし，式（3.1.19）の P^{head} と P^{tail} は，固有表現抽出と共参照解析の正解データではなく予測データを用いる場合がある．BRAN を改良した提案モデルの全体図を図 3.2 に示す．

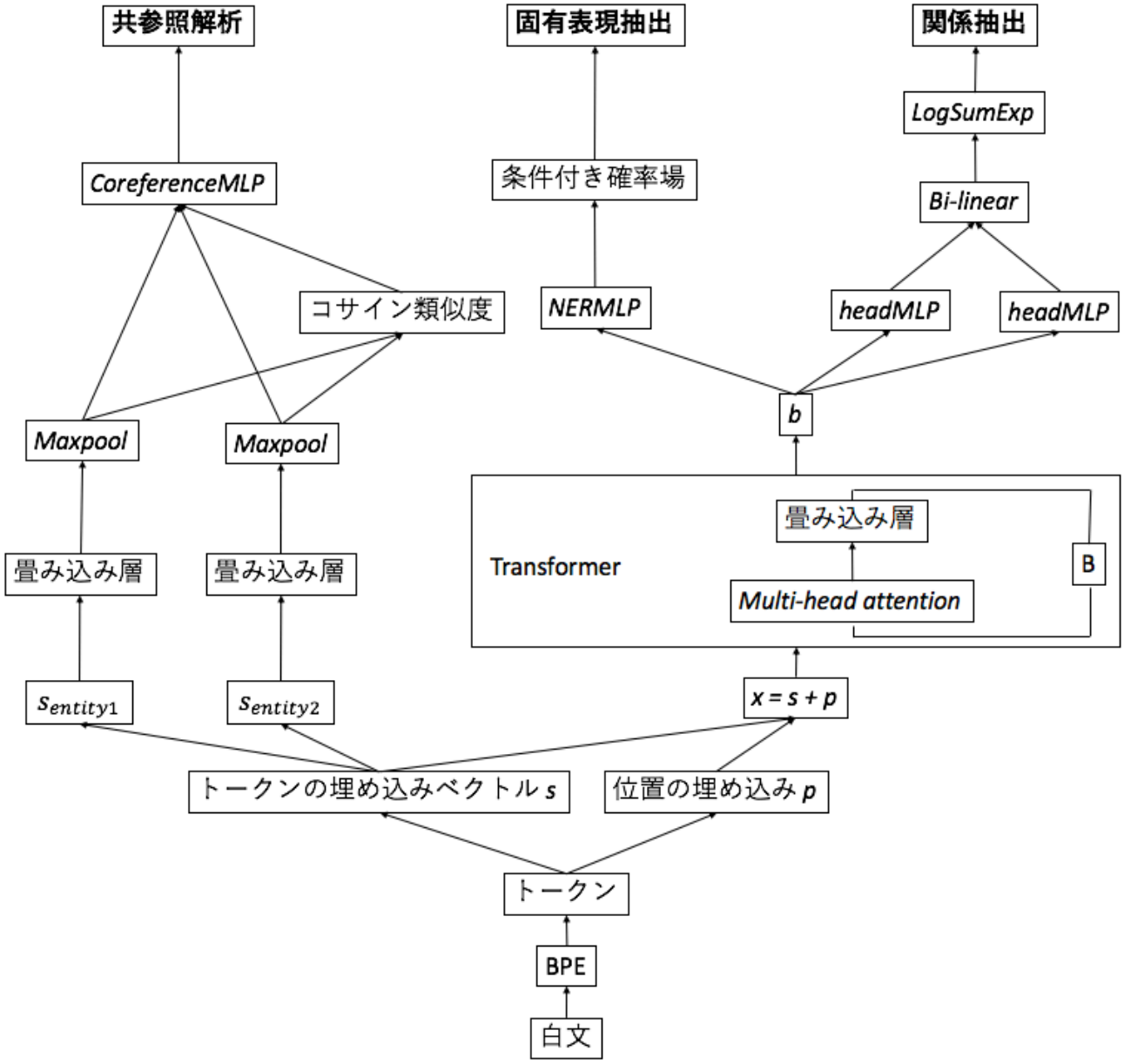
共参照解析
固有表現抽出
関係抽出
LogSumExp
CoreferenceMLP
条件付き確率場
Bi-linear
コサイン類似度
NERMLP
headMLP
headMLP
b
Maxpool
Maxpool
畳み込み層
Transformer
畳み込み層
B
畳み込み層
Multi-head attention
$s_{entity1}$
$s_{entity2}$
x = s + p
トークンの埋め込みベクトル s
位置の埋め込み p
トークン
BPE
白文

図 3.1　提案手法

第 4 章　実験

本章では実験に用いたデータ，評価方法，モデルなどのパラメータの設定と実験結果について説明する．

4.1　実験設定

4.1.1　データ

本研究で用いたデータは CDR である．CDR のアノテーションデータの一例を図 4.1 に示す．

```
439781|t|Indomethacin induced hypotension in sodium and volume depleted rats.
439781|a|After a single oral dose of 4 mg/kg indomethacin (IDM) to sodium and volume
depleted rats plasma renin activity (PRA) and systolic blood pressure fell significantly
within four hours. In sodium repleted animals indomethacin did not change systolic blood
pressure (BP) although plasma renin activity was decreased. Thus, indomethacin by
inhibition of prostaglandin synthesis may diminish the blood pressure maintaining effect
of the stimulated renin-angiotensin system in sodium and volume depletion.
439781  0       12      Indomethacin    Chemical        D007213
439781  21      32      hypotension     Disease D007022
439781  36      42      sodium  Chemical        D012964
439781  105     117     indomethacin    Chemical        D007213
439781  119     122     IDM     Chemical        D007213
439781  127     133     sodium  Chemical        D012964
439781  256     262     sodium  Chemical        D012964
439781  280     292     indomethacin    Chemical        D007213
439781  389     401     indomethacin    Chemical        D007213
439781  419     432     prostaglandin   Chemical        D011453
439781  518     529     angiotensin     Chemical        D000809
439781  540     546     sodium  Chemical        D012964
439781  CID     D007213 D007022
```

図 4.1　CDR の一例

図の一行目は，論文の ID，タイトルの印である「|t|」，論文のタイトルのテキストの順になっており，図の二行目は，論文の ID，概要文の印である「|a|」，論文の概要文のテキストの順になっている．次に，固有表現のラベル，関係のラベルの順に表記している．固有表現のラベルは「Chemical」，「Disease」の二種がある．関係のラベルは，「CID」がある．また図のように，固有表現のアノテーションは文頭からの文字数で位置がわかるようになっており，その表記，固有表現のラベルと固有表現ごとの ID がつけられている．本研究では，ID を用いて任意の固有表現ペアに

対し共参照のラベル，つまりペアになった固有表現が同一であるかないかの二値ラベルを作成する．図の最後の行は関係ラベルで，CID という関係ラベルを持つことを示し，固有表現の ID のペアが「Chemical」,「Disease」の順に表記されている．

CDR には 1500 の論文のタイトルと概要文がある．固有表現のラベルは,「Chemical」が 15935,「Disease」が 12850 ある．共参照ラベルは,「Chemical」の固有表現ペアの正例が 172069,「Disease」の固有表現ペアの正例が 98709,「Chemical」の固有表現ペアの負例が 340344,「Disease」の固有表現ペアの負例が 209976 ある．関係ラベルは，正例が 3116，負例が 12686 ある．CDR では，1500 の論文のデータを訓練データ，バリデーションデータ，テストデータに 500 ずつ分かれているが，本研究では無作為にバリデーションデータのうち 350 データを訓練データにし，訓練データ 850，バリデーションデータ 150，テストデータ 500 に分ける．

4.1.2 評価

固有表現抽出の評価には，BIO 方式を用いる．固有表現抽出で予測するラベルは,「B-Chemical」,「I-Chemical」,「B-Disease」,「I-Disease」,「O」の五種である．図 4.1 の固有表現「IDM」のラベルと概要文の「IDM」の出現箇所を見ると分かるように，固有表現はスペース区切りではなく，固有表現の前や後に文字である場合があることがわかる．そのため例えば，BPE により「(IDM)」が「(I」,「DM)」に分割された時，それぞれ固有表現ラベルとして「B-Chemical」,「I-Chemical」を付与する．また，次の節「実験結果」で Precision，Recall，F-measure で，既存研究と比較する．固有表現のラベル別に固有表現それぞれのトークンの範囲（文頭から何トークン目から何トークン目までか）を予測ラベル，正解ラベル共に求める．予測ラベル「Chemical」の固有表現それぞれののトークンの範囲の集合を *NER_Chemical_Predict*，正解ラベル「Chemical」の固有表現それぞれののトークンの範囲の集合を *NER_Chemical_Label* とする．*NER_Chemical_Predict* の要素の個数を *NER_Chemical_Predict_Num*，*NER_Chemical_Label* の要素の個数を *NER_Chemical_Label_Num* とする．また，*NER_Chemical_Predict* と *NER_Chemical_Label* の積集合の要素の個数（正解数）を *NER_Chemical_Correct_Num* とする．Precision，Recall，F-measure

の計算方法は，以下のような式で求める．

$$Precision_{NER} = \frac{NER\_Chemical\_Correct\_Num}{NER\_Chemical\_Precision\_Num} \tag{4.1.1}$$

$$Recall_{NER} = \frac{NER\_Chemical\_Correct\_Num}{NER\_Chemical\_Label\_Num} \tag{4.1.2}$$

$$F\text{-}measure_{NER} = \frac{2Recall_{NER} \cdot Precision_{NER}}{Recall_{NER} + Precision_{NER}} \tag{4.1.3}$$

これは，固有表現抽出ラベル「Disease」も同様に計算する．

共参照解析は，固有表現の正解データを用いて評価する．同じ文書中に含まれている同じ固有表現ラベルの付いている全てのトークンペアに対して共参照解析を行う．同じ ID の固有表現ペアを正例，異なる ID の固有表現ペアを負例とする．予測ラベルが正例の固有表現ペアの集合を *Coreference_Predict*，正解ラベルが正例の固有表現ペアの集合を *Coreference_Label*，*Coreference_Predict* の要素数を *Coreference_Predict_Num*，*Coreference_Label* の要素数を *Coreference_Label_Num* とする．また，*Coreference_Predict* と *Coreference_Label* の積集合の要素の個数（正解数）を *Coreference_Correct_Num* とする．Precision，Recall，F-measure の計算方法は，以下のような式で求める．

$$Precision_{Coreference} = \frac{Coreference\_Correct\_Num}{Coreference\_Precision\_Num} \tag{4.1.4}$$

$$Recall_{Coreference} = \frac{Coreference\_Correct\_Num}{Coreference\_Label\_Num} \tag{4.1.5}$$

$$F\text{-}measure_{Coreference} = \frac{2Recall_{Coreference} \cdot Precision_{Coreference}}{Recall_{Coreference} + Precision_{Coreference}} \tag{4.1.6}$$

関係抽出で予測するラベルは，「Chemical-induces Disease（CID）」「NULL（関係がない）」の二種である．関係抽出の評価は，固有表現抽出，共参照解析の正解データを用いる場合と，固有表現抽出，共参照解析の予測データを用いる場合がある．固有表現抽出，共参照解析の正解データを用いる場合二通りの評価方法がある．

- 固有表現抽出，共参照解析の正解データを用いた固有表現の ID ペアでの評価
  - 文書中に含まれている同じ ID の「Chemical」のラベルの付いている固有表現と同じ ID の「Disease」のラベルの付いている固有表現の全てのペアに対して関係ラベルのスコアを計算し，ID ペアに対して関係ラベルのスコアを計算する．
- 固有表現抽出，共参照解析の予測データを用いた固有表現ペアでの評価
  - 固有表現抽出を行う．
  -「Chemical」のラベルの付いている固有表現と「Disease」のラベルの付いている固有表現，それぞれ同タイプの全ての固有表現に対して共参照解析を行う．
  -「Chemical」の予測ラベルの付いている固有表現と「Disease」の予測ラベルの付いている固有表現の全てのペアに対して共参照解析の情報を用いて，関係抽出を行う．

固有表現抽出，共参照解析の予測データを用いる場合，先述の二つ目の方法をとる．以下の図 4.2，図 4.3 で図 4.1 のデータに対しての二つの関係抽出の評価手法の処理手順を示す．ただし，表記上データの一部を用いて示す．

図 4.2　関係抽出の評価手法 1

図 4.3　関係抽出の評価手法 2

予測ラベルが「CID」のペアの集合を *Relation_Predict*，正解ラベルが「CID」のペアの集合を *Relation_Label*，*Relation_Predict* の要素数を *Relation_Predict_Num*，*Relation_Label* の要素数を *Relation_Label_Num* とする．また，*Relation_Predict* と *Relation_Label* の積集合の要素の個数（正解数）を *Relation_Correct_Num* とする．Precision，Recall，F-measure の計算方法は，以下のような式で求める．

$$Precision_{RE} = \frac{Relation\_Correct\_Num}{Relation\_Precision\_Num} \tag{4.1.7}$$

$$Recall_{RE} = \frac{Relation\_Correct\_Num}{Relation\_Label\_Num} \tag{4.1.8}$$

$$F\text{-}measure_{RE} = \frac{2Recall_{RE} \cdot Precision_{RE}}{Recall_{RE} + Precision_{RE}} \tag{4.1.9}$$

4.1.3 パラメータ設定

モデルのパラメータを以下の表に示す．

表 4.1 モデルのパラメータ

埋め込みベクトルの次元	64
シーケンスの最大長	1111
B	2
H	4
トークンを unk トークンに変換する率	0.15
トークンのドロップアウト率	0.05
o ドロップアウト率	0.05
A ドロップアウト率	0.65

モデルのパラメータは BRAN と同じである．

次に訓練時の詳細な設定について説明する．訓練時，固有表現抽出と共参照解析，関係抽出を同確率で無作為に選択し学習する．共参照解析では，さらに同確率で「Chemical」，「Disease」のデータに対して学習する．関係抽出では，さらに同確率で「CID」，「NULL」のデータに対して学習する．また，バッチサイズは先行研究では固有表現抽出をおこなう場合バッチサイズ 16，関係抽出を行う場合バッチサイズ 32 としている．本研究の提案モデルの学習では，メモリ量により同じバッチサイズで学習することができなかったので，バッチサイズの比を合わせて，固有表現抽出，共参照解析を行う場合，バッチサイズは 8，関係抽出を行う場合，バッチサイズは 16 とする．最適化アルゴリズムは，Adam を用いている．

表 4.2　Adam のパラメータ

beta1	0.9
beta2	0.9
epsilon	0.0001
gradient clip	10

また，関係抽出を学習する際，勾配ノイズ $\eta = 0.1$ とする．また，一つのミニバッチデータを学習することを 1step と呼ぶ．Adam のパラメータ *alpha* を $alpha = alpha \times 0.75^{\frac{\text{step}}{25000}}$ で更新する．$\frac{15000}{\text{関係抽出のミニバッチサイズ}}$step（938step）ごとにバリデーションに移る．バリデーションデータを用いて，関係抽出の閾値を決める．関係抽出の「CID」のスコアに対し，閾値よりもスコアが大きければ予測ラベルを「CID」小さければ「NULL」とする．閾値は「0.5，0.6，0.7，0.75，0.8，0.85，0.9，0.95，0.975」を全て試し，最適な閾値を決める．バリデーションにおいての性能が向上したときテストに移る．テストの際の関係抽出の閾値は，バリデーションデータに対する最適な閾値を用いる．

4.2　実験結果

本節では，様々な点で BRAN と提案モデルの比較，提案モデルと従来手法を比較する．

まず，固有表現抽出結果を以下の表で比較する．

表 4.3　固有表現（Chemical）抽出結果の比較

手法	Precision	Recall	F-measure
BRAN	85.29	84.67	85.29
提案手法	91.17	89.35	90.52
提案手法 +Data	95.77	93.45	94.59

表 4.4　固有表現（Disease）抽出結果の比較

手法	Precision	Recall	F-measure
BRAN	68.93	72.88	70.85
提案手法	80.84	75.58	78.12
提案手法 +Data	85.30	84.88	85.09

ただし，「+Data」は固有表現抽出の CTD データセット [58] を追加したことを示す．

次に共参照解析の結果を示す．

表 4.5　提案手法の共参照解析結果

対象ラベル	Precision	Recall	F-measure
Chemical	97.43	94.49	95.94
Disease	96.46	89.29	92.73

次に以下の表で関係抽出の結果を比較する．表 4.6 では，評価手法ごとの BRAN の結果を比較するまた，表 4.7 の BRAN，提案手法は評価手法 2 を用いた結果を示し，表 4.7 の「()」は（）内のタスクについて予測ラベルを用いたことを示す．

表 4.6　BRAN による関係抽出結果の評価手法の比較

評価手法	Precision	Recall	F-measure
評価手法 1	56.99	74.20	64.47
評価手法 2	58.30	86.96	69.80

表 4.7　関係抽出結果の比較

手法	Precision	Recall	F-measure
BRAN	58.30	86.96	69.80
BRAN（固有表現抽出）	32.73	42.49	36.98
提案手法（固有表現抽出，共参照解析）	39.60	56.44	46.54
提案手法（固有表現抽出 +Data，共参照解析）	39.91	64.19	49.22
Zhou[38]	55.56	68.39	61.31
Zhou（固有表現抽出，共参照解析）[38]	42.59	49.91	45.96
Xu（固有表現抽出，共参照解析）[39]	55.67	58.4	57.0
Pons（固有表現抽出，共参照解析）[40]	51.34	53.85	52.56
Lowe（固有表現抽出，共参照解析）[41]	52.62	51.78	52.20

第 5 章　考察

本章では，実験結果を元に提案手法と比較手法について考察する．

まず，固有表現について考察する．表 4.3 と表 4.4 から，条件付き確率場を追加した結果，精度が向上し，CTD データを追加するとさらに精度が向上した．比較手法と比べても，高い精度となった．また，「提案手法 +Data」の精度は，CDR のデータの作成者の意見一致を考慮すると，これ以上の大幅な向上は難しいと考えられる．CDR のデータの作成者の意見一致を以下の表に示す．

表 5.1　アノテータの同意

データセット	Chemical	Disease
Train	95.2	86.0
Development	95.7	87.4
Test	96.3	88.8
All	96.1	87.5

上記の表のように，CDR のデータの固有表現はある程度意見の別れるアノテーションとなっている．表 5.1 から，本研究で提案した固有表現抽出のモデルは十分な精度を出していると考えられる．

次に，共参照解析結果について考察する．表 4.5 より，提案手法の共参照解析は本研究での共参照解析の設定において精度が高くなった．本研究での共参照解析の問題では，正例に同じ表現の固有表現ペアが多く，容易な問題設定であったことがわかる．また，共参照解析をモデルに組み込む事で，他のタスクと同時に学習する事で，他のタスクの影響し精度が向上すると考えていたが，変化は見られなかった．

次に，関係抽出結果について考察する．

表 4.6 の結果から，評価手法の比較を行う．評価手法 1（固有表現ペアを ID ごとに評価する）の方が評価手法 2（固有表現ペアごとに評価する）よりも精度が低くなった．これは，文書にある固有表現が多く存在すれば，それだけ関係の情報を文脈から取れるためであると考えられる．文書にある固有表現が多く存在すれば，その固有表現を含む関係ラベルが多くなり，より簡単に正解できるデータが多くなっ

たと考えられる．

表 4.7 の結果から，まず BRAN と提案手法を比較する．BRAN には，共参照解析のモデルがついていないので，正解データを使わない場合，固有表現の ID ごとに関係を抽出することができず，固有表現のペアごとに関係を推定しなければならない．その場合，BRAN では文脈の情報を効率よく扱かえず，共参照解析のモデルを追加した提案手法の方が精度が高くなった．

表 4.7 の結果から，提案手法と比較手法を比較する．

比較手法の Zhou[38] は機械学習を用いた手法で，構文解析などの情報を用いて単文に含まれる固有表現ペアに対して関係抽出を行う．比較手法の Zhou[38] と比較すると，外部知識を用いていない提案手法の方が精度が高く，文をまたぐ関係も抽出できるという点で，提案手法の方が優れていると言える．

比較手法の Xu[39] と Pons[40] はサポートベクターマシーンを用いた手法で，Xu[39] は MeSH，CTD，MEDication Indication Resource (MEDI)，Side Effect Resource (SIDER) といった外部知識を，Pons[40] は UniProt, UMLS, Medline[52] などのデータから得られたグラフベースの情報や，固有表現が何度出現したかなどの統計的特徴や Stanford CoreNLP parser による構文木の情報を用いて関係抽出を行う．比較手法の Xu[39] と Pons[40] の精度は大規模な外部知識を用いたものなので一概に比較はできないが，提案手法の方が精度が低くなった．手法としては，サポートベクターマシーンを用いた簡易な手法なので，外部知識などによる特徴量の効果が大きかったと考えられる．提案手法でも，これらの外部知識や構文木の情報を用い，特徴量として与えることで精度が向上すると考えられる．

比較手法の Lowe[41] はルールベースの手法である．比較手法の Lowe[41] と提案手法では，比較手法のほうが精度が高くなった．ただし，比較手法の Lowe[41] は固有表現抽出，共参照解析では辞書ベースの手法を，関係抽出ではルールベースの手法を用いており，提案手法の方が新たな問題には適応できると考えられる．

第 6 章　結論

本論文では，医学生物学論文のタイトルと概要文に対しての固有表現抽出や共参照解析，関係抽出の手法を提案した．BRAN をベースにし，固有表現抽出では条件付き確率場を追加し，共参照解析のモデルを追加し，固有表現抽出と共参照解析の予測を用いて関係抽出を行えるよう改良した．改良により，固有表現抽出のアノテーションが存在しない文書に対してより高精度で関係抽出を行えるようになった．また，データを追加し固有表現抽出の精度を向上した．本論文で提案したモデルは，機械学習を用いているので柔軟に問題に適応させることができる．また，文をまたいで関係を抽出でき，外部知識を利用しない場合手法の中では高い精度を出した．

今後のタスクとしては，大きく分けて三つ考えられる．一つ目は，外部知識と用いることである．外部知識を構造化し，特徴量としてモデルに組み込むことでさらなる精度向上が期待される．二つ目は，関係抽出を行うモデルの計算コストを減らすことである．式 3.1.18 式 3.1.19 の部分は，P^{head} や P^{tail} に含まれない A の要素は計算する必要はない．ただし，テンソルの計算上，計算コストを減らす操作を行うと処理速度が遅くなるという問題点がある．三つ目は，このモデルを他の分野に適用することである．他の分野に適応することで，モデル自体の特色や問題点がより浮き彫りになり，改善することができる．

謝辞

本研究を行うにあたり，多くの方々のご協力とご支援をいただきましたことを心より感謝いたします．主指導教員である松本裕治教授には，定例会や情報抽出勉強会において，研究に関するご指導，ご鞭撻いただきました．副指導教員である新保仁准教授には，定例会での進捗報告の際に，有益なアドバイスをいただきました．副指導教員である進藤裕之助教には，研究テーマの課題設定だけではなく，手法を選択するモチベーションや実験設定など，様々な点で御指導，ご鞭撻いただきました．北川裕子秘書をはじめ，研究生活を支えてくださった研究室の皆様に深く感謝を申し上げます．松本研で過ごした二年間は、多くの知識や技術を学び，物事を広い視野で捉え，深く考察できるようになり，とても実りのある研究生活を送れました．お世話になった皆様に心から感謝いたします．

参考文献

[1] Lev Ratinov and Dan roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155. Association got Computational Linguistics, 2009.

[2] 竹元義美，福島俊一，山田洋志．辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出．情報処理学会論文誌 Vol.42 No.6, pp.1580-1591, 2001.

[3] 内元清貴，馬青，村田真樹，小作浩美，内山将夫，井佐原均．最大エントロピーモデルと書き換え規則に基づく固有表現．自然言語処理 Vol.7 No.2, pp.63-90, 2000.

[4] Christophe D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, Vol.999. MIT Press, 1999.

[5] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *the 18th International Conference on Machine Learning 2001*, pp. 282–289, 2001.

[6] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[7] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Machine Learning Research*, Vol. 12, pp. 2493–2537, 2011.

[8] James Hammerton. Named entity recognition with long short-term memory. In *the seventh conference on Natural language learning at the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2003-Volume 4*, pp. 172–175. Association for Computational Linguistics, 2003.

[9] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Lin-*

guistics, Vol. 4, pp. 357–370, 2016.

[10] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint *arXiv:1603.01354*, 2016.

[11] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 97–102, 2017.

[12] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[13] Philip Gage. A new algorithm for data compression. *The C Users Journal 12(2)*, pp. 23-38, 1994.

[14] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford' s multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Conference on Natural Language Learning (CoNLL) Shared Task*, 2011.

[15] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 2001.

[16] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Association of Computational Linguistics (ACL)*, pp. 104– 111, 2002.

[17] Anders Björkelund and Richràd Farkas. Data-driven multilingual coreference resolution using resolver stacking. In *the Joint Conference on EMNLP and CoNLL: Shared Task*, pp. 49–55, 2012.

[18] Pascal Dennis and Jason Baldridge. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 660–669, 2008.

[19] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Meth-*

ods in Natural Language Processing, pp. 1971–1982, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[20] Kevin Clark and Christpher Manning. Entity-centric coreference resolution with model stacking. In *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1405–1415, 2015.

[21] Sam Wiseman, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Association of Computational Linguistics (ACL)*, pp. 92–100, 2015.

[22] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. In *Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL)*, 2016.

[23] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL ' 05, pp. 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[24] Mihai Surdeanu and Massimiliano Ciaramita. Robust information extraction with perceptrons. In *In NIST ACE*, 2007.

[25] Guodong Zhou, Min Zhang, Donghong Ji, and Qiaoming Zhu. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 728–736, 2007.

[26] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI' 07, pp. 2670–2676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[27] Kathrin Eichler, Holmer Hemsen, and Gnter Neumann. Unsupervised relation extraction from web documents. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. ELRA, 2008.

[28] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 113–120, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[29] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: Volume 2 - Volume 2, ACL ' 09, pp. 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[30] Alexander A. Morgan, Lynette Hirschman, Marc E. Colosimo, Alexander S. Yeh, and Jeffrey B. Colombe. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 396–410, 2004.

[31] Mark Craven and Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 77–86. AAAI Press, 1999.

[32] Patric Verga, Emma Strubell, Andrew McCallum. Simultaneously self-attention to all mentions for full-abstract biological relation extraction. In *HumanLanguage Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint *arXiv:1706.03762*, 2017.

[34] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint *arXiv:1607.06450*, 2016.

[35] Mikolov T., Sutskever I., Chen K. et al. Distributed representations of words and phrases and their compositionality. In *Proceedings of in Advances in Neural Information Processing Systems, USA*. pp. 3111–3119, 2013

[36] Leaman R., Wei C.-H., Lu Z.Y. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform., 7, S3, 2015.

[37] Leaman R., Doğan R.I., Lu Z.Y. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics, 29, 2909–2917, 2013.

[38] Huiwei Zhou, Huijie Deng, Long Chen, Yunlong Yang, Chen Jia, and Degen Huang. Exoloiting syntactic and semantics information for chemical-disease relation extraction. *Database* 2016.

[39] Xu J., Wu Y., Zhang Y. et al. UTH-CCB@ BioCreative V CDR task: identifying chemical-induced disease relations in biomedical text. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Spain* pp. 254–259, 2015.

[40] Pons E., Becker B.F.H., Akhondi S.A. et al. RELigator: chemical-disease relation extraction using prior knowledge and textual information. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Spain* pp. 247–253, 2015.

[41] Lowe D.M., O' Boyle N.M., Sayle R.A. LeadMine: disease identification and concept mapping using Wikipedia. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Spain* pp. 240–246, 2016.

[42] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 30–35, Baltimore, Maryland, June 2014. Association for Com-

putational Linguistics.

[43] Jiao Li, Yueoung Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus : a resource for chemical disease relation extraction. *Database*, 2016.

[44] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD ' 08, pp. 1247–1250, New York, NY, USA, 2008. ACM.

[45] SAIC. Proc. 7th message understanding conference (muc-7). 1998.

[46] IREX 実行委員会 (編). IREX ワークショップ予稿集. 1999.

[47] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pp. 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[48] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ace) program tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). ACL Anthology Identifier: L04-1011.

[49] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chain. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566, 1998.

[50] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational

Linguistics.

[51] Leaman R, Wei CH, Lu Z. NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem. Fourth BioCreative Challenge Evaluation; Bethesda, Maryland, USA. 2013. pp. 34–41.

[52] Kilicoglu H., Shin D., Fiszman M. et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28, 3158–3160, 2012.

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint *arXiv:1508.07909*, 2015.

[54] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. CNN-based ranking for biomedical entity normalization. *BMC Bioinf., 18, 385*, 2017.

[55] Wei-Oi Wei, Robert M Cronin, Hua Xu, Thomas A Lasko, Lisa Bastarache, and Joshua C Denny. Development and evaluation of an ensemble resource linking medications to their indications. *J. Am. Med. Inf. Assoc., 20, 954–961.* 2013.

[56] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol. 2010; 6():343,* 2010.

[57] Wei C.H., Peng Y., Leaman R. et al. Overview of the BioCreative V chemical disease relation (CDR) task. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. pp. 154–166. 2015

[58] Allan Peter Davis, Cynthia G. Murphy, Cynthia A. Saraceni-Richards, Michael C. Rosenstein, Thomas C. Wiegers, and Carolyn J. Mattingly. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res., 37, D786–D792. 2009.