修士論文

専門分野の辞書を用いた条件付き確率場再重み付け

辰巳 守祐
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 松本 裕治 教授
自然言語処理学研究室（情報科学領域）

令和 2 年 3 月 13 日

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

辰巳 守祐

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　(副指導教員)

専門分野の辞書を用いた条件付き確率場再重み付け*

辰巳 守祐

内容梗概

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，固有表現認識の学習データを自動生成できる Distant Supervision が注目されている．これらのモデルはマッチングしなかった全トークンに一般用語ラベルを付与する．しかし，一般的に，辞書が全ての固有表現を網羅するのは困難であるため，マッチングしなかったトークンを無視することは False Negative ラベルを生む問題がある．本稿では，このような False Negative を含むアノテーションデータを「不完全データ」と呼ぶ．

「不完全データ」の固有表現認識に関する従来研究の多くは，ラベルが 1 つに定まらないトークンに対して，可能性のあるラベル候補全てをトークンに付与し，ラベル候補の確率分布を学習する．特に，本研究のベースモデルは，ラベル候補の最適な確率分布を k 分割交差検証で求め，従来研究を上回る性能を出した．しかし，依然としてベースモデルには，3 つの課題がある．

1 つ目の課題は，学習初期段階で生じたノイズが学習後半まで伝播する問題である．これは，系列ラベリングの学習とノイズ緩和処理を同一プロセス，かつ同一情報資源で行なっていることが原因と考えられる．2 つ目の課題は，Recall と Precision のトレードオフ問題である．不完全データでの学習において，ベースモデルは従来手法と比較して，Recall は大幅に向上し，完全データで学習した従来手法の Recall に迫る高い値を出している．しかし，Precision に関しては，従来手法よりも低下している．3 つ目の課題は，専門用語の単語集合のうち，評価データの対象とする固有表現集合は一部に過ぎない問題である．ベースモデルの False Positive は，文脈情報と文字情報の観点では専門用語だが，評価データの対象とする固有表現ではない場合が多い．この問題を上記の素性のみで解くことは困難である．

提案手法では，辞書と不完全データを有効活用し，ベースモデルの Recall を保持しつつ，Precision の改善を試みた．実験結果より，提案手法は Recall を保持し

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 2 年 3 月 13 日.

つつ，Precision を改善でき，さらには，ゴールドデータで学習した場合のモデル性能に迫る認識性能を示せた．

キーワード

自然言語処理，系列ラベリング，固有表現認識, 専門分野辞書, 条件付き確率場

Reweighting in Conditional Random Fields using Expert-Domain Dictionary*

Shusuke Tatsumi

Abstract

Distant Supervision is a method that can automatically generate learning data for named entity recognition without manual annotation by dictionary matching. In these methods, we assign a generic term label to all unmatched tokens. However, in general, it is difficult for a dictionary to cover all named entities, so ignoring unmatched tokens has the problem of producing False Negative labels. In this paper, we call such annotation data including False Negative 'Incomplete data'.

Many previous studies on entity recognition of incomplete data assign all possible label candidates to a token whose label is not fixed to one. Then, the model is trained to predict the probability distribution of label candidates. In particular, the base model of this study obtained the optimal probability distribution of label candidates by k-cross validation, and outperformed the previous study. However, there are still three problems with the base model.

The first problem is noise propagation caused. The base model performs sequence labeling learning and noise reduction at the same time. This causes the noise at the beginning of learning to propagate to the latter half of learning.

The second problem is the trade-off between Recall and Precision. For learning with incomplete data, the recall of the base model was significantly improved compared to the conventional method. Furthermore, the base model showed high performance comparable to Recall of the conventional method trained on Gold data. However, in Precision, the base model was worse than the conventional method.

The third problem is that among the word sets of technical terms, the set of named entities targeted for evaluation data is only a part of them. The False Positive of the

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 13, 2020.

baseline is a technical term in terms of contextual information and character information, but is often not a named entity for evaluation data. It is difficult to discriminate these only with context information and character information.

We try to improve Precision by effectively using dictionaries and incomplete data without reducing Recall of the base model. From the experimental results, our proposed method can improve Precision without reducing Recall, and showed recognition performance comparable to that of the model trained with gold data.

Keywords:

Natural Language Processing, Sequence Labeling, Named Entity Recognition, Domain-Specific Dictionary, Conditional Random Fields

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

固有表現認識（Named Entity Recognition, NER）は，人名，場所，組織，薬品，時間，臨床手順，生体タンパク質などの固有表現をテキスト中から抽出するタスクである．近年，人手による精巧な素性選択を必要としない NER モデルが提案されてきたが，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．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．

そこで，アノテーションコストを削減するため，Distant Supervision による教師データの自動生成手法が注目されている．特に，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．Distant supervision では，専門用語辞書に含まれる固有表現をテキストコーパス中から検索し，NER モデルを構築する際の学習データ (正例) として用いる．一方，マッチしなかったコーパス中のトークンはすべて「一般用語」ラベル (負例に相当する) が付与される．Distant supervision では，人手のアノテーションを用いず学習データを自動生成できる利点があるが，一般的に，辞書が全ての固有表現を網羅するのは困難であるため，マッチングしなかった固有表現に一般用語ラベルを付与することにより，False Negative を生む可能性がある．本稿では，このような False Negative を含むアノテーションデータを「不完全データ」と呼ぶ．「不完全データ」の NER に関する従来研究の多く（Bellare ら [24], Carlson ら [25], Greenberg ら [26]）は，ラベルが 1 つに定まらないトークンに対して，マルチラベル（可能性のあるラベル候補全て）をトークンに付与し，ラベル候補の確率分布を学習する．特に，本研究のベースモデル（Jie ら [32]）は，マルチラベルの最適な確率分布を k 分割交差検証で求め，従来研究を上回る性能を出した．しかし，依然としてベースモデルには，次のような課題がある．

ベースモデルの課題

課題 1 系列ラベリングの学習とノイズ緩和処理を同一プロセス，かつ同一情報資源で行なっており，学習初期段階で生じたノイズが学習後半まで伝播する問題がある．別プロセス，かつ別情報資源からのノイズ緩和が必要．

課題 2 不完全データでの学習において，ベースモデルは従来手法と比較して，Recall は大幅に向上し，完全データで学習した従来手法の Recall に迫る高い値を出している．しかし，Precision に関しては，従来手法よりも低下している．

課題 3 ベースモデルの False Positive は，文脈情報と文字情報の観点では専門用語だが，評価データの対象とする固有表現ではない場合が多い．この問題の原因は，専門用語には様々な単語集合があるが，評価データの対象とする固有表現はその一部に過ぎない場合があるためと考えられる．

提案手法の目的

ベースモデルの課題に対し，本研究では以下のような対応策を提案する．

課題 1 への対応策 重み付け処理を改良する．具体的には,「ベースモデルの重み付け処理」と「辞書情報を用いた再重み付け処理」をパイプライン処理で行う．別プロセスとして「エンティティ・リンキング，知識グラフ埋め込みと弱教師ありクラスタリング」，別情報資源として「階層構造のある辞書」を使用し，ノイズ伝播の緩和を試みる．

課題 2 への対応策 本研究では，可能なかぎり Recall を低下させずに，False Positive を減少させ，Precision を向上させる．なお，Recall の向上は本研究の目標ではない．

課題 3 への対応策 辞書に含まれる様々な単語集合のうち，どの単語集合が評価データの対象とする固有表現の単語集合であるかを「辞書の階層構造情報」と「不完全データ」に基づき，事前に識別しておく．そして，NER モデル評価時，評価データの単語を辞書にリンキングし，リンク先の単語が属する単語集合が評価データの対象となる単語集合か否かを識別する．これにより，評価データの対象外の固有表現を誤って抽出するのを阻止でき，False Positive 減少が期待できる．

本実験を通して示せたこと

提案手法では，ベースモデルの Recall を保持しつつ，Precision 改善を試みた．実験結果より，提案手法は Recall を保持しつつ，Precision を改善でき，さらには，ゴールドデータで学習した場合のモデル性能に迫る認識性能を示すことができた．

1.2 本論文の貢献および得られた知見

- 辞書情報による，条件付き確率場の再重み付け手法を提案した．
- 提案手法が 3 データセット中，2 データセットで，ベースラインを上回った．
- 提案手法は「リンキング情報に基づく重み付け」と「辞書階層構造情報に基づく重み付け」の 2 つの要素技術から構成される．
- 「リンキング情報に基づく重み付け」は有効だったが,「辞書階層構造情報に基づく重み付け」の効果は限定的だった．

1.3 構成

本論文の構成は以下のとおりである．2 章ではまず，本研究の前提として，固有表現認識タスクの概要を説明し，不完全データにおける固有表現認識の関連研究を挙げる．3 章では，ベースモデルの問題点，提案手法の目的とモデル構成を説明をする．4 章で実験とその結果を示し，5 章で提案手法の挙動について分析を行う．最後に，6 章にて結論を述べる．

第2章　関連研究

2.1　固有表現認識

固有表現認識（Named Entity Recognition, NER）は，人名，場所，組織，薬品，時間，臨床手順，生体タンパク質などの固有表現をテキスト中から抽出するタスクである．具体的には，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) が用いられる．以下の表に例文「... lower lindane dose ... on acute renal failure ...」に対する系列ラベリングによる固有表現認識の例を示す．図 2.1 の例では，固有表現のクラスとして

Sentence: ...	lower	lindane	dose	...	on	acute	renal	failure	...
BIO: ...	O	B-CHE	O	...	O	B-DIS	I-DIS	I-DIS	...
BIOES: ...	O	S-CHE	O	...	O	B-DIS	I-DIS	E-DIS	...

図 2.1: 系列ラベリングによる固有表現認識

「CHE: Chemical」,「DIS: 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．

NER に関する研究は，Grishman ら [1] をはじめに，多くの研究がある（Sang ら [2], Sang ら [3], Piskorski ら [4], Segura-Bedmar ら [5], Bossy ら [6], Uzuner ら [7]）．

初期の NER モデルは，人手によるルールベース，辞書，正字法の特徴，オントロジーに基づいて識別していた．これらに続き，素性エンジニアリングと機械学習に基づいた NER モデルが考案された（Nadeau ら [8]）．

Collobert ら [9] を始めとする，ニューラルネットワークベースの NER モデルは最小限の素性エンジニアリングしか必要としないため，注目を集めている．これらのモデルは，通常，辞書やオントロジーなどのドメイン固有のリソースを必要

とせず，ドメイン依存が低い．様々なニューラルネットワークベースのモデルが提案されているが，その多くは，時系列データ処理に優れているリカレントニューラルネットワーク（Recurrent Neural Network，RNN）に基づいている．

2.2 専門分野への応用と課題

近年，人手による精巧な素性選択を必要としない NER モデルが提案されてきた（Li ら [10], Liu ら [11], Ma ら [12], Lample ら [13]）．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（Liu ら [11], Ma ら [12], Lample ら [13], Finkel ら [14]）．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 による教師データの自動生成手法が注目されており，フレーズマイニング（Shang ら [15]）, 固有表現認識（Ren ら [16], Fries ら [17], Ren ら [18]），Aspect Term Extraction（Giannakopoulos ら [19]），関係抽出（Mintz ら [20]）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ができる．従来の Distantly Supervised 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（Ren ら [18], Mintz ら [20]）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エンティティの範囲同定を行う手法（Ren ら [16], Fries ら [17]）などが提案されている．これらのモデルはマッチングしなかった全トークンに一般用語ラベルを付与する．しかし，一般的に，辞書が全ての固有表現を網羅するのは困難であるため，マッチングしなかった固有表現に一般用語ラベルを付与することにより，False Negative を生む可能性がある．本稿では，このような False Negative を含むアノテーションデータを「不完全データ」と呼ぶ．

2.3 不完全データのノイズ緩和策

「不完全データ」の NER に関する先行研究の多くは，CRF（Lafferty ら [21]），Structured Perceptron（Collins ら [22]），Max-Margin（Tsochantaridis ら [23]）モデルに基づいている．Bellare ら [24] は Missing Label Linear-Chain CRF，Carlson

ら [25] は Partial CRF，Greenberg ら [26] は EM Marginal CRF を提案している．いずれも本質的には，Latent Variable CRF（Quattoni ら [27]）と言える．また，このモデルは品詞タグ付やセグメンテーションタスク（Tsuboi ら [28], Liu ら [29], Yang ら [30]）にも使用されている．また，Yang ら [31] はこのモデルを強化学習と協調的に活用し，中国語の NER モデルを改善した．Greenberg ら [26] はこのモデルを生命科学分野での NER に適応し，有効性を示した．

これらの先行研究では，不完全アノテーションにより，トークンがマルチラベルを持つ場合，マルチラベルの確率分布を全ラベル均一に扱ってきた．しかし，Jie ら [32] はマルチラベルの最適な確率分布を k 分割交差検証で求める手法を提案し，先行研究を上回る性能を出した．

第3章　提案手法

3.1　ベースモデル

固有表現認識は，単語系列 X を入力とし，BIOES などのラベル系列 y を出力する系列ラベリングである．この系列ラベリングモデルの一つに Linear-Chain CRF（Lafferty ら [21]）がある．このモデルは完全なラベル付きデータ（シングルラベル系列　図 3.1，3.2）で，式 3.3 の損失関数を最小化するように学習する．ただし，$i = [1, ..., n]$ とする．

$$Input\ words = (X_1, X_2, ..., X_n) \tag{3.1}$$

$$Output\ labels = (y_1, y_2, ..., y_n) \tag{3.2}$$

$$L(w) = -\sum_i log\, p_w(y_i \mid X_i) \tag{3.3}$$

次に，不完全データのラベル系列 $y_i^{possible}$ の処理方法について考える．Jie ら [32] に従って，マルチラベル（付与される可能性のある全ラベル候補）を列挙し，このラベルセットを $C\left(y_i^{possible}\right)$ とする．上記の損失関数は式 3.4 となる．

$$L(w) = -\sum_i log \sum_{y \in C\left(y_i^{possible}\right)} q_D(y \mid X_i)\ p_w(y \mid X_i) \tag{3.4}$$

いくつかの関連手法を図3.1，3.2，3.3，3.4 に示す．この例では，固有表現 ‘acute’，‘renal’，‘failure’ がアノテーションされていない．図 3.1 はゴールドデータの場合のシングルラベルパスを示している．図 3.2 はアノテーションし損なった全トークンに ‘O’ ラベルを付与した，シンプルなモデルである．このモデルは，本質的には，式 3.4 における確率分布 q が任意の誤ったシングルラベルに全スコアを置いた場合と言える．

ここで，確率分布 q の意義について，先行研究（Bellare ら [24]）との比較を通して述べる．この手法では，アノテーションし損なったラベルを潜在変数と考え，

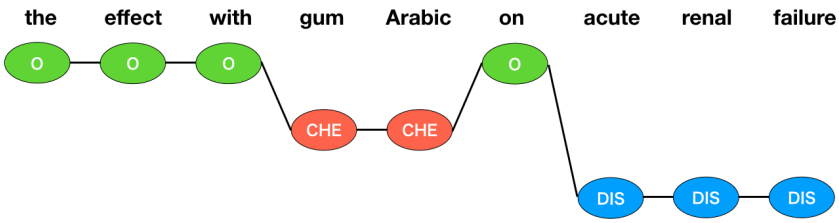

図 3.1: Gold

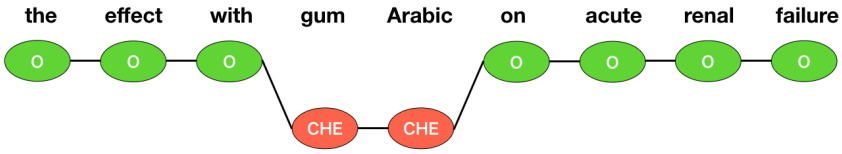

図 3.2: Simple

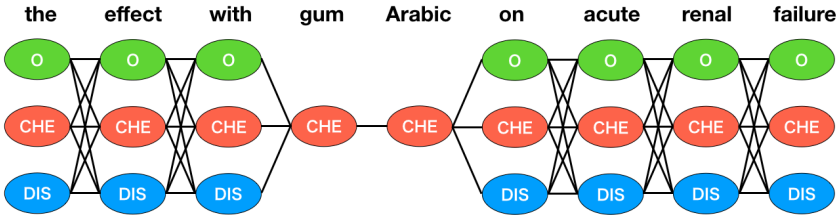

図 3.3: Uniform

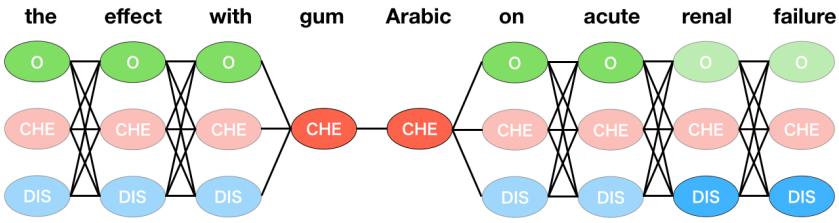

図 3.4: Distribution

式 3.5 の Latent Variable CRF で学習する．

$$L(w) = -\sum_{i} log \sum_{y \in C\left(y_i^{possible}\right)} p_w(y \mid X_i) \tag{3.5}$$

このモデルは Missing Label Linear-Chain CRF と呼ばれる．この手法は，図 3.3 で示すように，本質的には，確率分布 q が一様な場合である．この時，確率分布 q は確率値を全マルチラベル $C\left(y_i^{possible}\right)$ に等しく割り当てる．

一方，Jie ら [32] の手法では，図 3.4 に示すように，ゴールドパスにより近いパスになるような最適な確率値を，q が全マルチラベルに割り当てる．ただし，図のラベルの濃淡は確率値の高低を表している．

確率分布 q の推定

Jie ら [32] に従い，単語 X_i に対するラベル y の確率分布 q は，k 分割交差検証で求める．まず，マルチラベル部分に特定のシングルラベルを割り当てる初期化処理を行う．初期化処理については実験の章で取り上げる．次に，$(k-1)$ 分割データで学習を行い，その学習済み系列ラベリングの予測モデルで，残りの分割データに対してビタビアルゴリズムによるラベル推定を行い，確率分布 q を求める．確率分布 q は式 3.6 とする．彼らの研究では，q の求め方を 2 手法（ソフトアプローチとハードアプローチ）提案している．ハードアプローチでは，q は 0, 1 の値を取る離散的な分布で，最も可能性の高い 1 つのラベルに 1，その他に 0 を割り振る．一方，ソフトアプローチでは，q は連続的な分布で，マルチラベルそれぞれに特定の確率値を割り振る．本研究では，実装が公開されているハードアプローチで q を推定する．ハードアプローチでは，$(k-1)$ 分割データでの学習後，残りの分割データに対してビタビアルゴリズムによる系列ラベリングを行う．ビタビアルゴリズムでは，X_i のラベルが y_j であるスコア P_{i,y_j} を推定する．本研究では，関連研究（Liu ら [33]，Ma ら [34]，Lample ら [13]）に従って，式 3.7 で予測ラベル系列のスコアを求める．スコア Φ はラベル y_i からラベル y_{i+1} への遷移可能性を表す．y_j は式 3.10 で示すように，固有表現の範囲ラベル $\{B, I, E, S\}$ と固有表現のクラスタイプ $\{Class\,Types\}$ を組み合わせて得られるラベルに一般用語のラベル ‘O’ を加えたラベル集合から構成される．

$$q_D(y \mid X_i) = \frac{e^{s(X_i, y)}}{\sum_{y \in C\left(y^{possible}\right)} e^{s(X_i, y)}} \tag{3.6}$$

$$s(X,y) = \sum_{i=0}^{n} \Phi_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i}, \tag{3.7}$$
$$\Phi_{y_i,y_{i+1}}, P_{i,y_i} \in R$$

$$\Phi_{y_i,y_{i+1}} : Transition\,Probability\,Score \tag{3.8}$$

$$P_{i,y_j} : Score\ for\ the\ word\ X_i\ being\ the\ label\ y_j \tag{3.9}$$

$$y_j \in [\{B,I,E,S\} \times \{Class\,Types\} + \{O\}] \tag{3.10}$$

3.2 提案手法の目的

ベースモデルの課題

ベースモデルの主な課題は以下の 3 点である.

課題 1 学習初期で生じたノイズが学習後半まで伝播する問題.

課題 2 Precision の低下.

課題 3 False Positive は，文脈情報と文字情報の観点では専門用語だが，評価データの対象とする固有表現ではない場合が多い.

提案手法の目的

ベースモデルの課題に対し，本研究では以下のような対応策を提案する.

課題 1 への対応策 重み付け処理を改良する．具体的には,「ベースモデルの重み付け処理（節 3.1）」と「辞書情報を用いた再重み付け処理（節 3.4）」をパイプライン処理で行う.

課題 2 への対応策 本研究では，False Positive の減少を試みる．辞書情報による識別（節 3.5, 3.6）により，一般用語である可能性が高いと判定された単語において，マルチラベルのうち ‘O’ ラベルに高い確率値を割り当てる．具体的な処理については，後述の節 3.4 の式 3.11, 3.12 で定義する.

課題 3 への対応策 専門用語のうち，評価データの対象とする単語集合を節 3.6 のように機械的に識別し，評価データの対象外の固有表現の誤抽出を防止する.

3.3 専門分野の辞書情報

本研究では，専門分野の辞書情報として「単語の文字情報」と「階層構造情報（ルートから各単語へのパス）」を使用する．

3.4 再重み付け

ベースモデルにおける，ビタビアルゴリズム推定の P_{i,y_j} を算出後，専門分野の辞書情報を用いて，再重み付けを行う．直感的には，辞書中の専門用語へマッチングできなかった単語を一般用語と仮定し，一般用語と判定された X_i のラベル y_i が 'O' と予測されやすくなるように重みを調整する．具体的には，ラベル y_i の全マルチラベルのうち，ラベル 'O' のスコアに超パラメータの実数値である α，または β を加算する．再重み付けは *Algorithm* 1 に示すように 2 行程から成る．

まず，X_i の辞書へのエンティティ・リンキング（節 3.5）を行う．リンキングできなかった X_i は一般用語と識別し，式 3.11 で再重み付けする．次に，リンキングできた X_i は辞書の階層構造情報に基づく識別（節 3.6）を行う．一般用語と識別されたもののみ，式 3.12 で再重み付けする．

Algorithm 1 Re-Weighting Overview

Require:

$Words = [X_0, X_1, ..., X_n]$ ▷ X_i *has* X_i.character *and* X_i.labels

procedure RE-WEIGHTING(*Words*)
  for all *Words* **do**
    if *Word*.label is Incomplete **then** ▷ *Word*.label ∈ [*Gold*, *Incomplete*]
      LinkingResult ⇐ ENTITYLINKING(*Word*.character) ▷ *LinkingResult* ∈ [*Linked*, *Unlinked*]
      if *LinkingResult* is Unlinked **then**
        WEIGHTING(*Word*.labels)
      else
        IdentifyingResult ⇐ IDENTIFYING(*Word*) ▷ *IdentifyingResult* ∈ [*NamedEntity*, *GeneralTerm*]
        if *IdentifyingResult* is General Term **then**
          WEIGHTING(*Word*.labels)

また，学習終盤まで再重み付けすると，辞書情報の影響を過度に受け，モデル予測の Recall が向上し難くなる可能性がある．そこで，学習序盤は，確証バイア

ス緩和のために強めに再重み付けし，学習終盤はモデル予測の Recall を向上させるために弱めに再重み付けする．具体的には，α, β を固定せず，式 3.13，3.14 に従って，学習の進行に伴い減少させる．η はハイパーパラメータ，$\alpha_{init}, \beta_{init}$ は α, β の初期値，*epoch_count* は現時点で終了したエポックの回数を示す．

$$P\left(X_i, y_{j=\text{`}O'}\right) = P\left(X_i, y_{j=\text{`}O'}\right) + \alpha, \quad P\left(X_i, y_{j=\text{`}O'}\right), \alpha \in R \tag{3.11}$$

$$P\left(X_i, y_{j=\text{`}O'}\right) = P\left(X_i, y_{j=\text{`}O'}\right) + \beta, \quad P\left(X_i, y_{j=\text{`}O'}\right), \beta \in R \tag{3.12}$$

$$\alpha = max\left(0, \alpha_{init} - \eta \cdot epoch\_count\right) \tag{3.13}$$

$$\beta = max\left(0, \beta_{init} - \eta \cdot epoch\_count\right) \tag{3.14}$$

3.5 辞書へのエンティティ・リンキング

エンティティ・リンキングでは素性に文脈情報を使用する手法が知られている．しかし，本研究で扱う専門用語は低頻度語句が多く，学習に必要な文脈を十分確保できない．よって，本研究では文脈情報は使用せず，単語の表層形のみを扱う．具体的には，レーベンシュタイン距離に基づいたエンティティ・リンキングを行う．レーベンシュタイン距離は，2 つの文字列がどの程度異なっているかを示す距離の一種であり，具体的には，1 文字の挿入・削除・置換によって，一方の文字列をもう一方の文字列に変形するのに必要な手順の最小回数として定義される．本研究のエンティティ・リンキングのアルゴリズムを *Algorithm* 2 に記す．X_i とのレーベンシュタイン距離が MIN_DISTANCE 以下の辞書単語のうち，距離が最短の単語を返す．MIN_DISTANCE 以下の単語が無ければ,「リンク先なし（No_Linked）」を返す．ただし，単語の長さが MIN_CHARACTER 未満の単語は上記の基準を満たすリンク先が多過ぎるため，リンキングは行わず,「リンク先なし（No_Linked）」を返す．

Algorithm 2 Entity Linking

Require:

Word : *Characters*

```
 1: procedure ENTITY LINKING(Word)
 2:     if Word.length < MIN_CHARACTER then
 3:         return No_Linked
 4:     else
 5:         for all NamedEntity ∈ Dictionary do
 6:             Distance = LEVENSHTEIN_DISTANCE(Word, NamedEntity)
 7:             if Distance < MIN_DISTANCE then
 8:                 Candidates.add(NamedEntity)
 9:             SelectedEntity ⇐ GETCLOSESTENTITY(Candidates)
10:         if SelectedEntity is Exist then
11:             return SelectedEntity
12:         else
13:             return No_Linked
```

3.6 辞書の階層構造情報に基づく識別

辞書の階層構造情報を用いて，単語 X_i が一般用語か固有表現か識別する．

3.6.1 準備段階

系列ラベリングモデルとは独立に，単語 X_i が一般用語か固有表現かを識別する．このため，事前に階層構造情報を含む辞書と「不完全データ」を用いた準備を行う．以下の 3 行程から成る．アルゴリズム を *Algorithm* 3 に記す．

Algorithm 3 Discrimination using Hierarchical Structure Information

Require:

1: $IncompAnnoNamedEntitySet \Leftarrow$ GETNAMEDENTITYSET($IncompAnnoData$) ▷ IncompAnnoData is Incomplete Annotation Data for NER
2: **procedure** PREPARATION($Word$)
3: $Embeddings \Leftarrow$ KNOWLEDGEGRAPHEMBEDDING($HierarchicalStructure$)
4: $Clusters \Leftarrow$ UNSPUERVISEDCLUSTERING($Embeddings$)
5: $LabeledClusters \Leftarrow$ CLASSIFYING($IncompAnnoNamedEntitySet$, $Clusers$)

6: **procedure** INFERENCE($Word$)
7: $LinkedWord \Leftarrow$ ENTITYLINKING($Word$)
8: $LabeledCluster \Leftarrow$ SEARCHINGCLUSTER($LinkedWord$)
9: **if** $LabeledCluster$.label is GeneralTerm **then**
10: **return** $Word\,is\,GeneralTerm$
11: **else**
12: **return** $Word\,is\,NamedEntity$

13: **function** CLASSIFYING($IncompAnnoNamedEntitySet$, $Clusers$)
14: **for all** $Cluster \in Clusters$ **do**
15: **for all** $Word \in Cluster$ **do**
16: **if** $Word \in IncompAnnoNamedEntitySet$ **then**
17: $count+ = 1$
18: $rate = count \div Cluster$.num_words
19: **if** $rate > \mu$ **then** ▷ μ is threshold. $\mu \in (0, 1)$
20: $Clusters$.label $\Leftarrow NamedEntity$
21: **else**
22: $Clusters$.label $\Leftarrow GeneralTerm$

第 1 行程 (図 3.5)

辞書の階層構造情報を知識グラフ埋め込みによりベクトル化する．これにより，辞書に含まれる固有表現をベクトルとして扱うことができる．

第 2 行程 (図 3.6)

全固有表現ベクトルをインスタンスとした，教師なしクラスタリングを行う．これにより，クラスタが得られる．

第 3 行程 (図 3.7)

全クラスタを「不完全データ」を用いて，一般用語クラスタと固有表現クラスタに分類する．具体的には，各クラスタにおいて,「不完全データ」の固有表現の割合が閾値 μ を超えれば固有表現クラスタ，超えなければ一般用語クラスタとする．

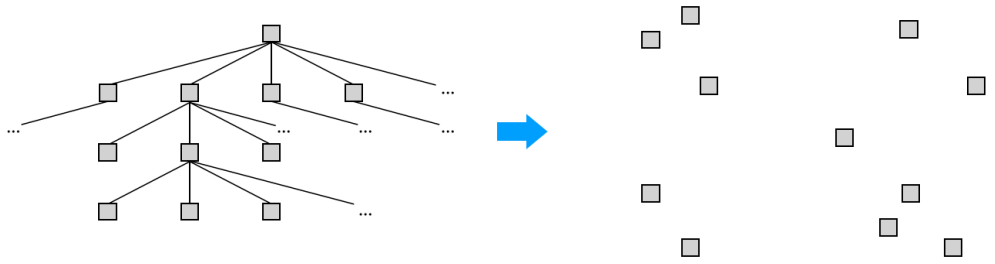

図 3.5: 第 1 行程

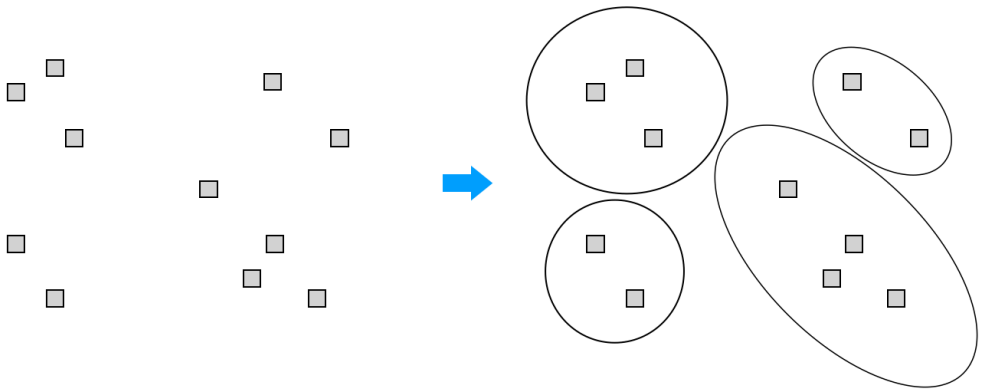

図 3.6: 第 2 行程

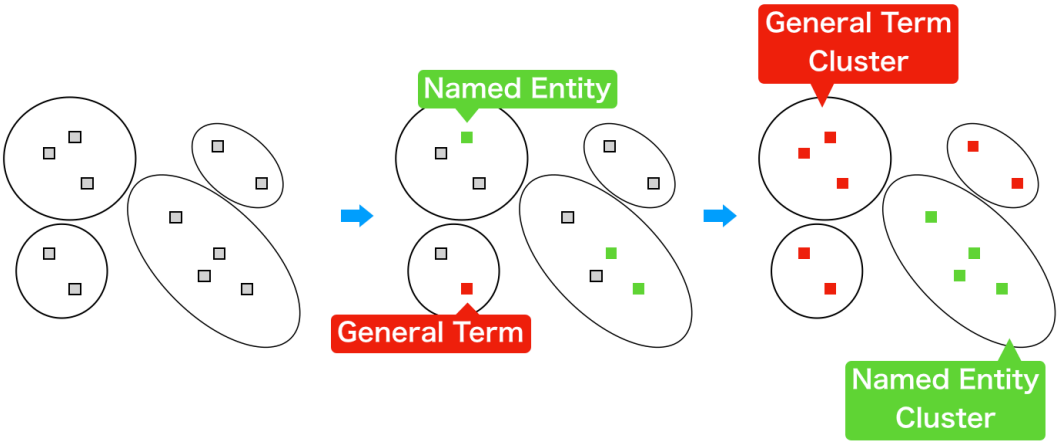

図 3.7: 第 3 行程

3.6.2 推論段階

単語 X_i を辞書にリンキングし，リンク先の単語が属するクラスタを求める．そして，そのクラスタが一般用語クラスタであれば，X_i を一般用語とし，固有表現クラスタであれば，X_i を固有表現とする．

3.7 モデルが学習しているパラメータ

確証バイアスの緩和が目的であるため,「系列ラベリング」と「提案部分」の学習は独立して行う．すなわち，NER の LSTM-CRF の学習時，提案部分のパラメータ学習はしない．NER の LSTM-CRF の学習はベースモデルと同様にする．提案部分において，全行程の内，NER の LSTM-CRF の学習と独立のパラメータ学習が可能なのは,「エンティティ・リンキング」,「クラスタリングのインスタンス作成のための，辞書のグラフ埋め込み」と「教師なしクラスタリング」部分である．ただ，エンティティ・リンキングについては，今回はルールベース手法（レーベンシュタイン距離に基づくリンキング手法）を採用した為，パラメータ学習しない．辞書のグラフ埋め込みと教師なしクラスタリングについては，先行研究の論文に基づき，パラメータ学習する．使用する先行研究については，実験の章で述べる．

第4章　実験

4.1　実験設定

本実験では，既存の固有表現認識モデルと提案手法の固有表現認識性能を比較し，提案手法の有効性を検証する．

4.1.1　データセット

評価データの一覧を表4.1に示す．全データセット共に無料公開されており，いずれも訓練データ，開発データ，評価データに分けられている．詳細は下記に記す．

BC5CDR （Li ら [35]）

正式名称は BioCreative V Chemical and Disease Mention Recognition task で，対象は医学用語と化学用語．PubMed から収集された論文 1,500 本の概要から成り，15,935 の Chemical Mentions と 12,852 の Disease Mentions を含む．

NCBI-Disease （Dogan ら [36]）

対象は医学用語．PubMed から収集された論文 793 本の概要から成り，6,892 の Disease Mentions を含む．

CHEMDNER （Krallinger ら [37]）

対象は化学用語．PubMed から収集された論文 10,000 本の概要から成り，25,610 の TRIVIAL, 19,138 の SYSTEMATIC, 13,118 の ABBREVIATION, 12,028 の FORMULA, 11,935 の FAMILY, 1,824 の IDENTIFIER, 589 の MULTIPLE, 113 の NO CLASS を含む．

CHEMDNER は対象分野が幅広く，これら全てを辞書で網羅するのは困難である．そこで，本研究では TRIVIAL，SYSTEMATIC，FAMILY のみを固有表現認識の評価対象とし，その他は一般用語として評価する．

表 4.1: データセット構成

	BC5CDR	NCBI-Disease	CHEMDNER
Domain	Biomedical	Biomedical	Biomedical
Entity Types	Disease, Chemical	Disease	Chemical
Articles	1,500	793	10,000
Mentions	15,935 Chemical 12,852 Disease	6,892 Disease	25,610 Trivial 19,138 Systematic 11,935 Family

不完全データ

不完全なアノテーションデータ（アノテーションされた固有表現の Precision は 100 だが，Recall は 100 ではない）しか入手できないという実験設定は，専門分野の固有表現認識において一般的である．同様の実験設定とするため，本実験では，ゴールドデータの固有表現のうち，ランダムに一定数の固有表現を欠落させ，それらのラベルを ‘O’ とする．正しくアノテーションされた固有表現の割合を固有表現網羅率 ρ で表す．例えば，ρ=0.6 の場合，全固有表現の 60%は保持するが，残りの 40%は除去し，それらに ‘O’ ラベルを付与する．

4.1.2 専門分野辞書

本研究では，生命科学分野の辞書である CTD[1] を使用する．具体的には，BC5CDR による評価実験では，CTD Chemical vocabularies と CTD Disease vocabularies を合わせた辞書を，NCBI-Disease では，CTD Disease vocabularies を，CHEMDNER では，CTD Chemical vocabularies を使用する．CTD Chemical vocabularies には 172,861，CTD Disease vocabularies には 12,984 の語句が収録されている．

[1] http://ctdbase.org/

4.1.3 評価指標

評価指標に F1 スコア (式 4.3) を使用する．Precision (式 4.1) と Recall (式 4.2) も併記する．

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.2}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4.3}$$

4.1.4 詳細設定

系列ラベリングモデルの訓練に *Stochastic Gradient Descent*（*SGD*）を使用する．バッチサイズ：10，モメンタム：0.0，学習率：0.01，ドロップアウト：0.5 に設定する．系列ラベリング学習のイテレーション回数は 10 エポック．各イテレーションにおける，k 分割交差検証のイテレーション回数は 10 エポックで，$k = 2$ とする．先行研究（Jie ら [32], Nivre ら [38]）によると，k をさらに大きくした場合も，無視できる程度の性能向上しか得られない．重みに加算する α，β は共に 5.0，η は 0.5 とする．単語表現としては，共起確率に基づく分散表現である 100 次元 GloVe，文字情報 RNN，文脈情報に基づく分散表現である ELMo を使用する．エンティティ・リンキングでは，レーベンシュタイン距離と単語長の超パラメータ（3.5 節）を $MIN\_DISTANCE = 5, MIN\_CHARACTER = 7$ とし，大文字小文字の区別はしない．辞書階層構造情報での識別では，閾値 $\mu = 0.001$ とし，知識グラフ埋め込みには TransE（Bordes ら [39]）を，教師なしクラスタリングには k-means 法（Pelleg ら [40]）を使用する．上記詳細設定の一覧を表 4.2, 4.3, 4.4 に記す．

4.1.5 単語表現の事前学習

GloVe（Pennington ら [41]）は生命科学関連の文献情報データベースである MEDLINE[2]から収集した論文概要で事前学習したものを使用する．ELMo（Gardner ら [42]）は各データセットの生データで事前学習したものを使用する．

[2]https://www.ncbi.nlm.nih.gov/pubmed/

表 4.2: 系列ラベリングモデル学習の詳細設定

変数	値
勾配法	*SGD*
バッチサイズ	10
モメンタム	0.0
学習率	0.01
ドロップアウト	0.5
系列ラベリング学習エポック数	10
1 分割あたりのエポック数	10
k	2
α	5.0
β	5.0
η	0.5
単語表現	$GloVe, CharRNN, ELMo$

表 4.3: エンティティ・リンキングの詳細設定

変数	値
$MIN\_DISTANCE$	5
$MIN\_CHARACTER$	7
大文字小文字の区別	なし

表 4.4: 階層構造情報による識別時の詳細設定

変数	値
μ	0.001
知識グラフ埋め込み	*TransE*
クラスタリング	$k-means$

4.2 比較手法

4.2.1 Linear-Chain BiLSTM-CRF

Linear-Chain BiLSTM-CRF（Lample ら [13] が提案したモデル）は全ての単語にシングルラベルを付与して処理する．欠落させた固有表現のラベルは 'O' とする．また，これとは別に，ゴールドデータで学習した場合の性能とも比較するため，ゴールドデータで学習したモデルの性能評価も行う．

4.2.2 Multi Label BiLSTM-CRF

Multi Label BiLSTM-CRF（Jie ら [32] が提案したモデル）は本研究のベースモデルであり，保持した固有表現にはシングルラベル，その他の単語にはマルチラベルを付与して処理する．

4.2.3 提案モデル

本研究では，Ours-Linking，Ours-Fixed，Ours-All の 3 つの手法を提案する．Ours-Linking は辞書階層構造情報を使用せず，リンキング情報のみで再重み付けを行なっている．Ours-Fixed $\alpha\beta$ では，学習の進行に伴う α, β の減少を行わず，常時固定している．Ours-All では，提案した全ての機能を使用している．まとめると，表 4.5 のようになる．

表 4.5: 提案モデルの機能比較

Model	Decreasing $\alpha\beta$	Entity Linking	Hierarchical Structure
Ours-Linking	○	○	×
Ours-Fixed $\alpha\beta$	×	○	○
Ours-All	○	○	○

なお，Multi Label BiLSTM-CRF と提案モデルの初期化では，Jie ら [32] に従った．具体的には，図 3.2 のように，それぞれの分割学習において，保持された固有表現は B, I, E, S，その他は O ラベルとして学習し，得られた結果をマルチラベル確率分布 q の初期化に利用する．

4.3 実験結果

4.3.1 固有表現認識性能の比較

固有表現網羅率 $\rho = 0.5$ での実験結果を，データセット別に，表 4.6〜表 4.8 に示す．まず，不完全データで学習したベースラインについては，Linear Chain BiLSTM-CRF は，Precision が高く，Recall が低い．Multi Label BiLSTM-CRF は，Precision が低く，Recall が高い．これらのベースラインと提案手法の Ours-All を比較すると，Linear Chain BiLSTM-CRF との比較では，Precision は低下するものの，Recall は大幅に向上した．結果的に，F1 で提案手法が上回った．また，Multi Label BiLSTM-CRF との比較では，Recall は低下するものの，Precision は大幅に向上した．結果的に，F1 において，3 データセット中，2 データセットで，提案手法が上回った．次に，提案手法の有効性を人手によるゴールドデータで学習した場合とも比較した．提案手法の実験設定ではアノテーションが不完全であるにも関わらず，ゴールドデータでの学習モデルに迫る結果が得られた．

以上より，不完全データにおける，固有表現認識の課題に対する提案手法の有効性を示せた．

4.3.2 学習進行に伴う $\alpha\beta$ 減少の効果

Ours-Fixed $\alpha\beta$ と Ours-All の比較より，学習進行に伴う $\alpha\beta$ 減少が有効だと分かった．Ours-Fixed $\alpha\beta$ は Precision が非常に高いが，Recall が低下し，それに伴い F1 も低い．これに対し，Ours-All は Precision は Ours-Fixed $\alpha\beta$ ほど高くはないが，Recall が高い．結果的に，Ours-Fixed $\alpha\beta$ より Ours-All の方が F1 が高かった．

4.3.3 辞書階層構造情報の効果

提案手法はリンキング情報と辞書階層構造情報の 2 つの情報を用いて識別している．そこで，これらの情報のいずれが識別へ大きく寄与しているのかを分析した．具体的には，リンキング情報のみを用いた Ours-Linking と辞書階層構造情報も使用した Ours-All を比較実験した．結果より，リンキング情報の効果は見られるが，辞書階層構造情報の効果は見られなかった．理論的には Ours-All の方が Ours-Linking よりも Precision が高くなる．しかし，得られた実験結果では Ours-Linking の方が高かく，3 データセット中，2 データセットで，Ours-Linking の方が F1 が優れていた．

表 4.6: BC5CDR

	Annotation	Precision	Recall	F1
Linear Chain BiLSTM-CRF	Gold	88.18	87.58	87.88
Linear Chain BiLSTM-CRF	Incomplete	78.79	74.63	76.65
Multi Label BiLSTM-CRF	Incomplete	74.03	85.51	79.35
Ours-Linking	Incomplete	81.06	83.80	82.41
Ours-Fixed α β	Incomplete	84.01	78.78	81.31
Ours-All	Incomplete	80.50	84.63	82.51

表 4.7: NCBI-Disease

	Annotation	Precision	Recall	F1
Linear Chain BiLSTM-CRF	Gold	81.95	82.29	82.12
Linear Chain BiLSTM-CRF	Incomplete	82.09	60.62	69.74
Multi Label BiLSTM-CRF	Incomplete	72.54	77.60	74.99
Ours-Linking	Incomplete	79.46	76.15	77.77
Ours-Fixed α β	Incomplete	80.48	63.13	70.75
Ours-All	Incomplete	77.42	76.77	77.09

表 4.8: CHEMDNER

	Annotation	Precision	Recall	F1
Linear Chain BiLSTM-CRF	Gold	89.37	85.57	87.43
Linear Chain BiLSTM-CRF	Incomplete	89.46	66.39	76.21
Multi Label BiLSTM-CRF	Incomplete	77.06	81.17	79.06
Ours-Linking	Incomplete	80.75	78.33	79.52
Ours-Fixed α β	Incomplete	83.47	72.87	77.82
Ours-All	Incomplete	79.63	78.21	78.91

第5章　分析

5.1　学習進行に伴う$\alpha\beta$減少とRecall向上の関係

Ours-Fixedでは$\alpha\beta$固定したのに対して，Ours-Allでは固定せず，学習の進行に伴い減少させ，モデル予測のRecallを向上させた．そこで，仮説通りに学習進行に伴ってRecallが向上しているかを調べるため，ベースモデルMulti Label BiLSTM-CRF，提案手法Ours-Fixed，提案手法Ours-Allの比較分析を行った．具体的には，3つの手法において，学習の進行に伴うPrecision, Recall, F1の変化を描画し，学習の進行とRecallの向上の相関について分析した．図5.1〜5.3はデータセットBC5CDRでの実験結果，図5.4〜5.6はデータセットNCBI-Diseaseでの実験結果，図5.7〜5.9はデータセットCHEMDNERでの実験結果を示す．

学習序盤では，提案手法Ours-Fixed，Ours-AllともにPrecisionでベースモデルを上回った．学習終盤においては，提案手法Ours-FixedはRecallがベースモデルと比較して低いが，提案手法Ours-Allは$\alpha\beta$の減少によりRecallが向上し，ベースモデルと同程度まで回復した．以上より，仮説通り，学習進行に伴ったRecall向上がなされていると分かった．

5.2　提案手法の効果と固有表現網羅率ρの関係

提案手法の効果とρの関係を調べるため，異なる$\rho = [0.1, 0.2, ..., 0.9]$で実験を行った．比較手法はベースモデルMulti Label BiLSTM-CRFと提案手法Ours-All，データセットはBC5CDRを使用した．結果を図5.10〜5.12に示す．Precisionの比較では，異なるρ全般的に両手法とも安定した値となっており，常に提案手法がベースモデルを上回った．Recallの比較では，両手法とも$\rho = 0.1$時点では値が低く，$\rho = 0.3$にかけて上昇し，$\rho = 0.4$以降は安定した値を示した．$\rho = 0.3$以下では，ベースモデルの方が高いRecallを示したが，$\rho = 0.4$以降では，両手法に大きな差は見られなかった．F1の比較では，$\rho = 0.3$以下では，Recall比較の影響を大きく受け，ベースモデルの方が高い値を示した．一方，$\rho = 0.4$以降では，Precision比較の提供を大きく受け，提案手法の方が高い値を示した．

以上より，コーパスの固有表現網羅率が一定水準より高い場合，提案手法が有効だと分かった．

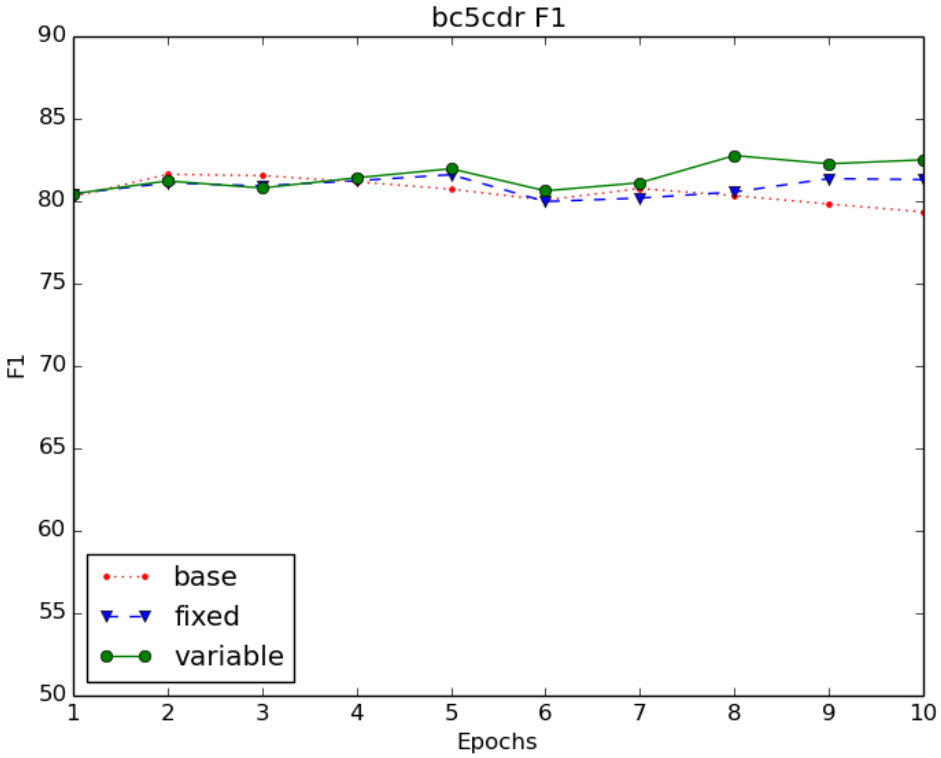

図 5.1: BC5CDR F1

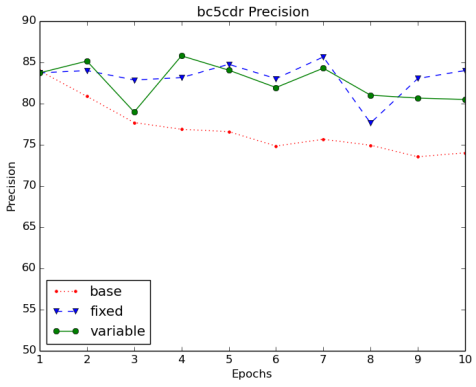

図 5.2: BC5CDR Precision

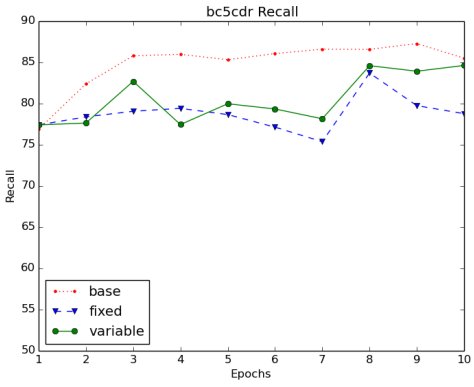

図 5.3: BC5CDR Recall

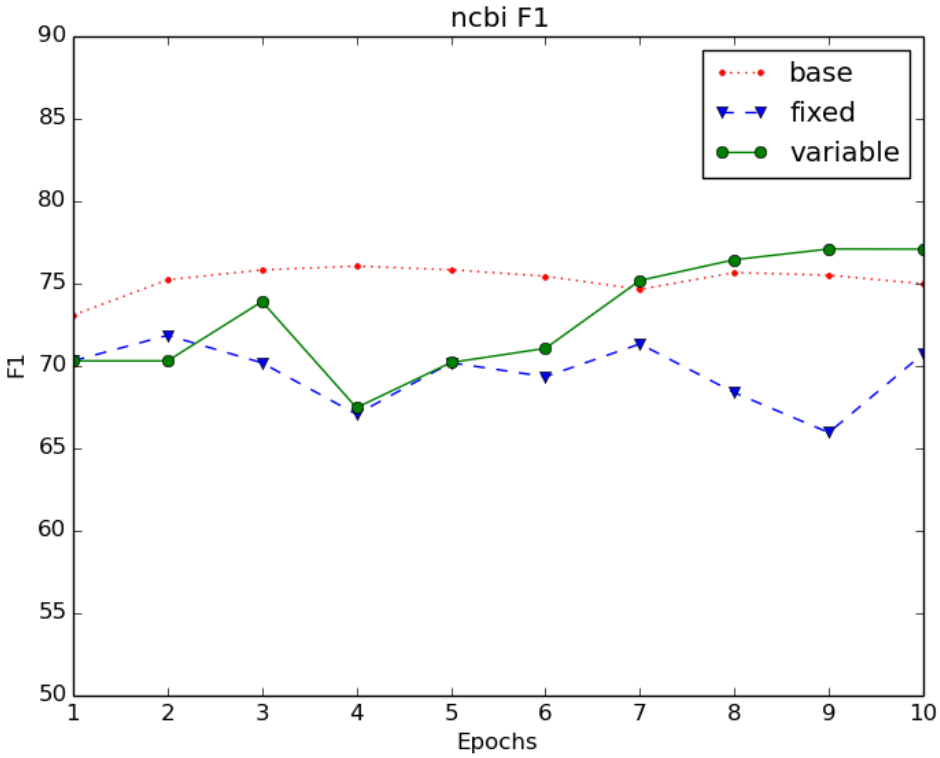

図 5.4: NCBI-Disease F1

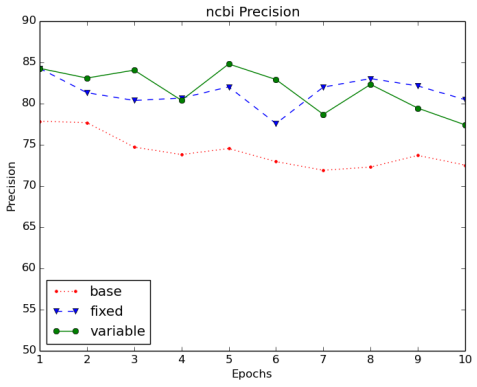

図 5.5: NCBI-Disease Precision

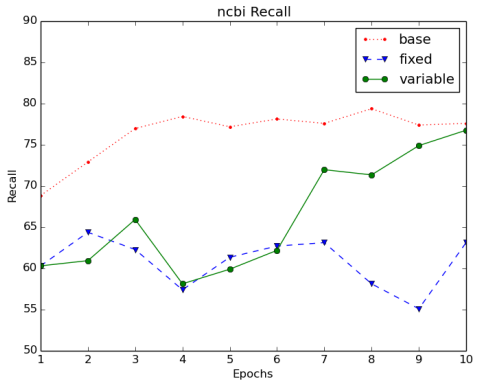

図 5.6: NCBI-Disease Recall

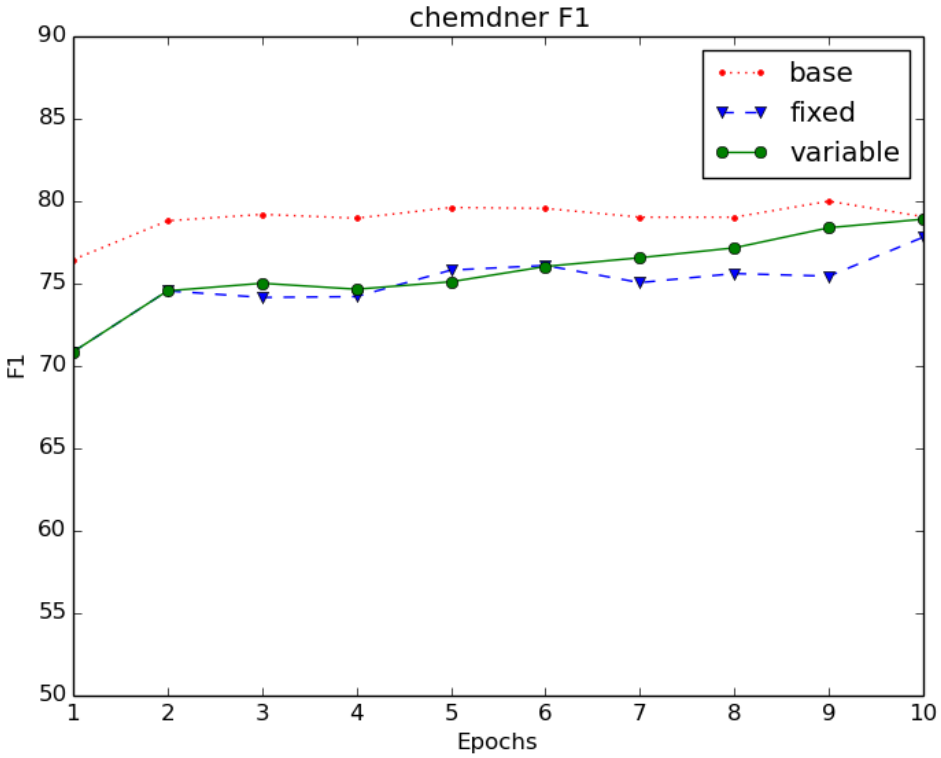

図 5.7: CHEMDNER F1

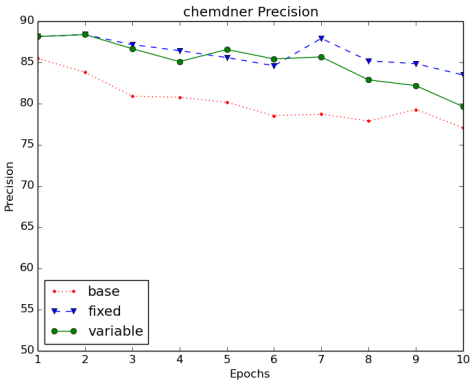

図 5.8: CHEMDNER Precision

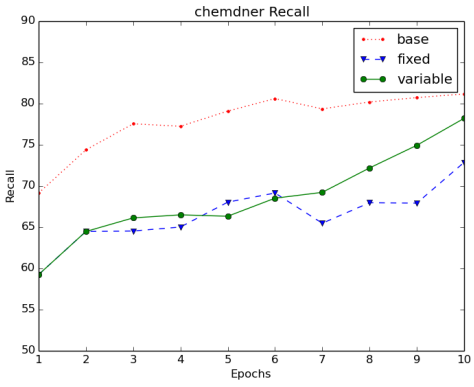

図 5.9: CHEMDNER Recall

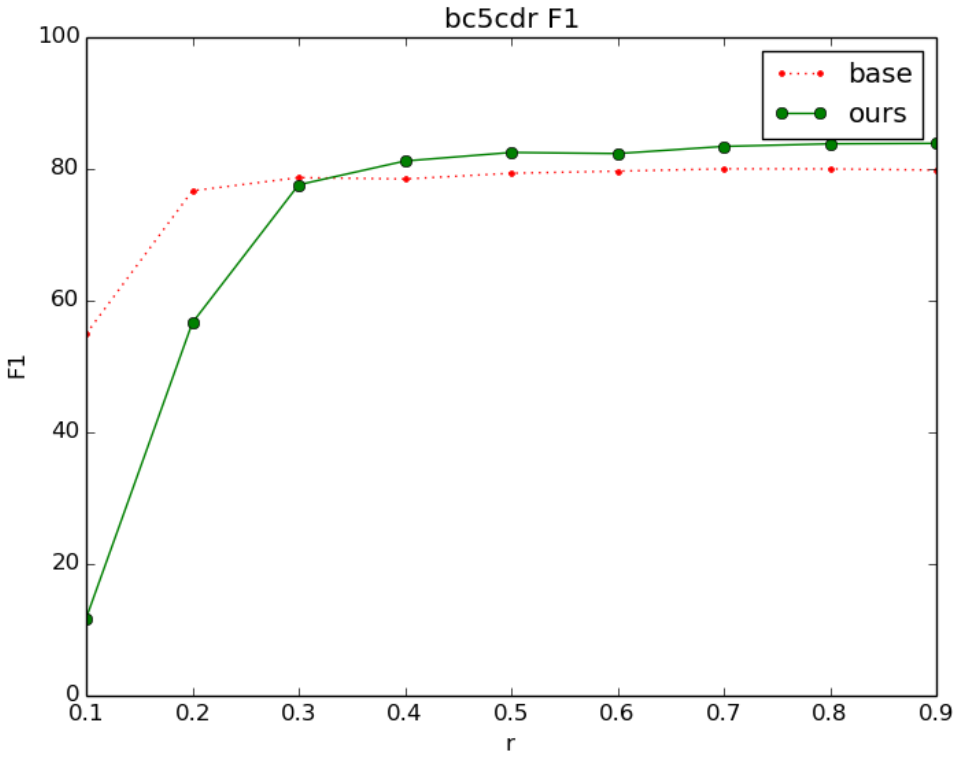

図 5.10: 固有表現網羅率 ρ と F1

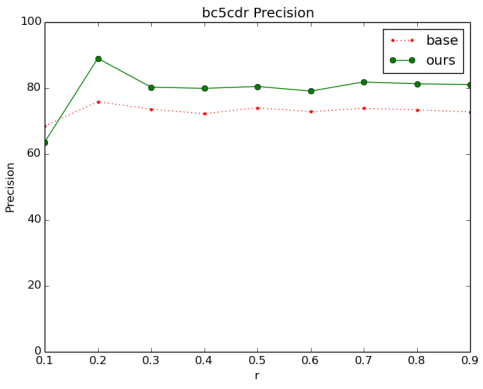

図 5.11: 固有表現網羅率 ρ と Precision

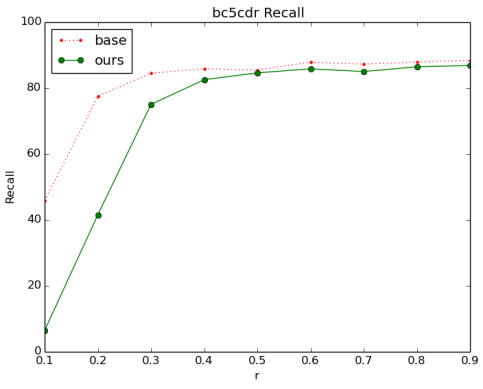

図 5.12: 固有表現網羅率 ρ と Recall

第6章　おわりに

6.1　結論

不完全データでの固有表現認識では，慣例的な手法はRecall向上が困難だった．一方，Recall向上を目指したベースモデルでは，Recall向上に有効だったものの，Precisionが低下する問題があった．そこで，提案手法では，ベースモデルのRecallを保持しつつ，Precisionの改善を試みた．実験結果より，提案手法はRecallを保持しつつ，Precisionを改善でき，さらには，ゴールドデータに迫る認識性能を示せた．以上より，不完全データにおける，固有表現認識の課題に対する提案手法の有効性を示せた．本研究の貢献および得られた知見を以下にまとめる．

- 辞書情報による，条件付き確率場の再重み付け手法を提案した．
- 提案手法が3データセット中，2データセットで，ベースラインを上回った．
- 提案手法は「リンキング情報に基づく重み付け」と「辞書階層構造情報に基づく重み付け」の2つの要素技術から構成される．
- 「リンキング情報に基づく重み付け」は有効だったが,「辞書階層構造情報に基づく重み付け」の効果は限定的だった．

6.2　展望

本研究の展望を以下にまとめる．

- リンキングのみのモデルの方が辞書階層構造情報を使用したモデルよりもPrecision高かったが，これは理論的に不自然であり，今後分析が必要である．
- 辞書による重み付け処理方法を改良する．現在,「エンティティ・リンキング」と「辞書階層構造情報による重み付け」は完全に分離されているが，今後はこれらの処理を協調的に行えるように改良する．

- 本研究では，単語区分が不明瞭である問題から，マルチワードをシングルワードに分割して擬似的にリンキングしており，本質的には，マルチワードを考慮できていない．しかし，これではマルチワードを取りこぼす問題がある．そこで，今後は，リンキング時に，単語分割タスクを並行して解くことで，マルチワードのリンキングを行う．

謝 辞

松本裕治教授，新保仁准教授，進藤裕之助教には研究会や勉強会で，中村哲教授には中間報告会で助言をいただいた．本学システム微生物学研究室の森浩禎教授と野村航研究員からは，生命科学分野の辞書に関する助言をいただいた．後藤啓介研究員と近藤修平研究員にはプログラム実装で，北川祐子秘書には，研究室での生活面でお世話になった．研究室では優秀な先輩，同期，後輩に恵まれた．

英ユニバーシティ・カレッジ・ロンドンへの研究留学では当研究テーマに取り組み，Professor Sebastian Riedel，Lecturer Pontus Stenetorp から助言をいただいた．留学資金は「文部科学省 トビタテ留学 JAPAN 日本代表プログラム」からご支援いただいた．お世話になった皆様に謝意を表する．

参考文献

[1] Ralph Grishman and Beth Sundheim. Message understanding conference- 6: A brief history. In *COLING*, 1996.

[2] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[3] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[4] Jakub Piskorski, Lidia Pivovarova, Jan Snajder, Josef Steinberger, and Roman Yangarber. The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. pp. 76–85, 01 2017.

[5] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[6] Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. BioNLP shared task 2013 – an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 161–169, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, Vol. 18, pp. 552–6, 06 2011.

[8] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, Vol. 30, , 01 2007.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, Vol. 12, pp. 2493–2537, November 2011.

[10] Jing li, Aixin Sun, Ray Han, and Chenliang Li. A survey on deep learning for named entity recognition, 12 2018.

[11] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. 09 2017.

[12] Xuezhe Ma and Eduard H. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, Vol. abs/1603.01354, , 2016.

[13] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[15] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, Vol. PP, , 02 2017.

[16] Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 995–1004, New York, NY, USA, 2015. ACM.

[17] Jason A. Fries, Sen Wu, Alexander Ratner, and Christopher Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *CoRR*, Vol. abs/1704.06360, , 2017.

[18] Xiang Ren, Ahmed El-Kishky, Chi Wang, and Jiawei Han. Automatic entity recognition and typing in massive text corpora. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pp. 1025–1028, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[19] Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. Unsupervised aspect term extraction with b-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 180–188, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[20] Mike Mintz, Steven Bills, Rion Snow, and Daniel 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*, pp. 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[21] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[22] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002*

Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 1–8. Association for Computational Linguistics, July 2002.

[23] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, Vol. 6, pp. 1453–1484, December 2005.

[24] Kedar Bellare and Andrew Mccallum. Learning extractors from unlabeled text using relevant databases. 01 2007.

[25] Andrew Carlson, Scott Gaffney, and Flavian Vasile. Learning a named entity tagger from gazetteers with the partial perceptron. pp. 7–13, 01 2009.

[26] Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2824–2829, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[27] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Conditional random fields for object recognition. 01 2004.

[28] Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. Training conditional random fields using incomplete annotations. pp. 897–904, 01 2008.

[29] Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. Domain adaptation for crf-based chinese word segmentation using free annotations. 10 2014.

[30] Fan Yang and Paul Vozila. Semi-supervised Chinese word segmentation using partial-label learning with conditional random fields. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 90–98, Doha, Qatar, October 2014. Association for Computational Linguistics.

[31] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2159–2169, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[32] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In *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)*, pp. 729–734, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[33] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. 09 2017.

[34] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[35] Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, Vol. 2016, p. baw068, 05 2016.

[36] Rezarta Dogan, Robert Leaman, and Zhiyong lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, Vol. 47, , 01 2014.

[37] 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 Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, 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 Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, No. 1, p. S2, 2015.

[38] Joakim Nivre and Ryan McDonald. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT*, pp. 950–958, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[39] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 2787–2795. Curran Associates, Inc., 2013.

[40] Dan Pelleg and Andrew Moore. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pp. 277–281, New York, NY, USA, 1999. ACM.

[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[42] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

発表文献

[1] 辰巳守祐, 進藤裕之, 松本裕治 化学ドメインにおける教師無し固有表現抽出, 言語処理学会第 *25* 回年次大会発表論文集, 2019 年 3 月.

[2] 辰巳守祐, 後藤啓介, 進藤裕之, 松本裕治 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, 情報処理学会 第 *241* 回自然言語処理研究会, 2019 年 8 月.

[3] 辰巳守祐, 野村航, 武藤愛, 森浩禎, 松本裕治 遺伝的相互作用データおよび文献情報の機械学習による相互作用ネットワーク推定, 第 *14* 回日本ゲノム微生物学会年会, 2020 年 3 月.

[4] 野村航, 辰巳守祐, 武藤愛, 森浩禎, 松本裕治 自然言語処理による知識抽出のための辞書構築と生物実験結果解釈への応用, 第 *14* 回日本ゲノム微生物学会年会, 2020 年 3 月.

[5] Shusuke Tatsumi, Pontus Stenetorp, Yuji Matsumoto *Reweighting in Conditional Random Fields using an Expert-Domain Dictionary,* 言語処理学会第 *26* 回年次大会発表論文集, 2020 年 3 月.