自然言語教示によるフレーズ抽出器の学習に関する研究

齊藤　亮将† 　小林　滉河† 　若林　啓 ††

† †† 筑波大学　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{s1711514,s1921631}@s.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　フレーズ抽出とは，テキストデータから特定の単語やフレーズを抽出する情報抽出タスクであり，様々な技術に利用されている重要なタスクである．機械学習を用いてフレーズ抽出器を学習するには，訓練データにアノテーションコーパスを使用するが，複雑なモデルになるほど必要となるデータ数も増加し，アノテーションタスクに伴うコスト面での問題点が挙げられる．そこでアノテーションコストと訓練データ数を削減し，低コストで効率的なフレーズ抽出器の学習を行うことが本研究の目的である．自然言語による説明文（自然言語教示）を用いて生成された訓練データを用いた，点予測によるモデル学習手法を提案し，実験結果によりそのモデルの性能を示した．

キーワード　自然言語処理，自然言語教示，フレーズ抽出，点予測，Labeling Function

1 はじめに

1.1 背　　景

フレーズ抽出とは，テキストデータ，すなわち自然言語文から特定の単語やフレーズを抽出する情報抽出タスクである．図 1 を用いて説明すると，テキストから人名を抽出することを目的としているのであれば「田中 太郎」という人名（フレーズ）を抽出し，テキストから固有表現を抽出することを目的としているのであれば「田中 太郎」と「千葉」を抽出するということである．

このフレーズ抽出は，様々な技術に利用されている．文章中から人名・地名・組織名といった固有表現の抽出を行う固有表現抽出（Named Entity Recognition; NER）[1], [2] もフレーズ抽出の一つで，質疑応答システムや対話システムなどに利用されている．また，専門用語を抽出することを目的としたフレーズ抽出である専門用語抽出 [3] は，専門用語を専門用語辞書のようなコンピュータ上で扱いやすい形へと変換するために利用されている．これは無数にある論文の中に，それぞれの著者によって名称が異なるような専門用語があり，それを統計的にまとめることなどに役立っている．

このように，テキストデータからフレーズを抽出することは重要なタスクであり，膨大なテキストデータを扱うにはフレーズ抽出の自動化が必須である．近年多くの手法では，フレーズ抽出を自動化させるためには機械学習を用いて，高精度な「フレーズ抽出器」を学習させる．フレーズ抽出器は，図 1 のようなフルアノテーションコーパスを訓練データとすることで学習することができる．

アノテーションコーパスとは，各テキストに対して，機械学習を行う上で使用される情報が付与されたテキストの集合で

文字列 x	田中	太郎	は	千葉	出身	だ	.
ラベル列 y	B-PER	I-PER	O	B-LOC	O	O	O

図 1　フルアノテーションコーパスの例

あり，この情報を付与する人をアノテータと呼ぶ．その中でも，各テキスト全てのトークンにアノテーションがされているコーパスをフルアノテーションコーパスと呼ぶ．情報抽出タスクにおいて付与する情報としてはいくつかの形式があるが，ここでは IOB2 タグ [4] と呼ばれる形式について説明する．IOB2 タグでは，抽出するカテゴリに属する単語に特定のタグを付与する（人名であれば PERSON を表す PER, 組織名であれば ORGANIZATION を表す ORG など）．また，固有表現の先頭トークンには追加で B タグを付与し，連続するそれ以降のトークンには I タグを付与する．そして，固有表現ではないトークンには O タグを付与する．図 1 は，このタグ付与を行ったフルアノテーションコーパスの例である．

固有表現抽出でよく用いられる機械学習モデルには, 条件付き確率場（Conditional Random Field; CRF）[1], [2] や Bidirectional Long Short-Term Memory（Bi-LSTM）と CRF を組み合わせた Bi-LSTM-CRF [5] などが挙げられる．

先にも述べたように，この手法でフレーズ抽出器を学習させるにはフルアノテーションコーパスが必要であり，このフルアノテーションコーパスを作成する中でいくつかの懸念点が挙げられる. まず一つは，アノテーション作業に時間やコストが非常にかかる点である．一般的に機械学習モデルの訓練データとして，膨大な量のコーパスが必要になることが知られており，その一つ一つにアノテーション作業を行うことを考えると時間やコストがかかることは明白である．また，必要なコーパスの量に比例して負担も大きくなるため，より多くのコーパスを必要とする複雑なモデルを扱う際には，大きな問題部分となることが考えられる．次に，アノテータが限定される場合についてである．抽出したいカテゴリが人名や組織名などであれば，多くの人間がアノテーションを行うことが出来る．しかし，バイオ分野や物質材料分野などの特定の分野における専門用語を抽出したいカテゴリとすると，専門的な知識のない人間がアノテーションを行うのは困難であり，その分野に精通している専門家をアノテータとしなくてはならない．また，外部に公開できないデータを扱う場合も考えなくてはならない．近年では，アノ

テーションコーパスを作成するのにクラウドソーシングを利用するケースも珍しくなく，必要なデータ量が膨大であればあるほどコストはかかってしまうが，人手を集める問題に関しては比較的容易に解決できる．しかし，外部に公開できないデータを扱ってモデル学習を行う場合は，クラウドソーシングのような手法も取れず，限定された人間でアノテーションを行わなければならない．以上からも分かるように，機械学習を用いて学習させようとした時に，大量のアノテーションコーパスを用意することが障壁となる場合が予想され，訓練データとして必要なアノテーションコーパスの量を削減したり，アノテーション作業 1 タスクあたりのコストを軽減させることが，フレーズ抽出器モデルの学習コスト軽減につながると考えられる．

また，Hancock ら [15] は，単純なラベル付与以上の情報を獲得するために各ラベル付与決定に対しての説明を自然言語文で表現し，それを分類器の学習のために使用したフレームワーク BabbleLabel を提案した．そして，この自然言語文での説明文を自然言語教示と呼んでいる．これは，一つの自然言語教示文から得た情報を Labeling Function の形に変換することで，複数の訓練データ作成に活用することが出来るため，アノテーションコスト削減に役立つと考えられる．

1.2 本研究の目的

本研究では，Hancock らの手法をフレーズ抽出のタスクに応用し，アノテーションコストと訓練データ数を削減することにより，低コストで効率的なフレーズ抽出器の学習を行う手法を提案する．従来のアノテーション方法では，アノテータの思考を間接的に訓練データのラベルに反映させている．しかしこれでは，ある単語 x がどうして抽出するラベル y であるのかというラベル付与の理由部分が不明である．この隠された理由部分に多くの情報が含まれていると考えられることから，それを活用するために自然言語教示を用いる．そして，その自然言語文を解析し，それをフレーズ抽出のための関数や単語辞書を検索する上でのキーワードとして使用する．最終的にそれらを使用して未使用の対象コーパス全体を機械的にラベル付けをすることで，擬似的に大量の訓練データを作成する．この手法により，必要な訓練データ数の削減とアノテータのコストを軽減することが可能になる．提案手法の有用性を示すために，フレーズ抽出タスク，固有表現抽出タスクでよく使用されている CoNLL-2003 English dataset [6] を使用してフレーズ抽出器の精度を検証した．

また，本研究での貢献としては以下の三つを挙げる．

- フレーズ抽出タスクにおいて，自然言語文（自然言語教示）を用いた枠組みの提案.
- Hnacock らの手法をフレーズ抽出タスクに応用した，独自の Labeling Function の提案.
- 実験を用いた自然言語教示のフレーズ抽出タスクに対する有効性の確認.

1.3 本論文の構成

本稿は次のとおりに構成される．第 2 章では，関連する研究について述べ．本研究の位置づけを明らかにする．第 3 章では，自然言語教示によるフレーズ抽出器の学習の提案手法について説明する．第 4 章では，既存手法と提案手法についての比較実験を行う．第 5 章では，結論と本研究の今後の展開について議論を行う．

2 先 行 研 究

2.1 フレーズ抽出

本節では，フレーズ抽出に関する研究について紹介する．1 章でも述べたようにフレーズ抽出とは，テキスト中から人名や地名を抽出する固有表現抽出や化学物質名，病名と言った専門性の高い専門用語を抽出する専門用語抽出など，想定される特定のフレーズを自然言語文から抽出するタスクを指し，それらは質疑応答システムや対話システム，情報抽出など多様に利用されている．

固有表現抽出は，系列ラベリングタスクとして解決されることが一般的であり，Morwal ら [1] や Ekbal ら [2] は，隠れ状態を持つマルコフ過程である隠れマルコフモデル（Hidden Markov Model; HMM）による性能を示した．その後，系列全体を考慮するために，Lafferty ら [7] や Das ら [8] は，系列データに対して確率論的モデルである条件付き確率場（Conditional Random Field; CRF）によるモデルの性能を示した．また，Riaz [9] や Alfred ら [10] は，手動で設定されたルールベースに基づく手法を示した，近年では，固有表現抽出に深層学習を用いることも多く，Lample ら [5] は，双方向 LSTM（Bidirectional Long Short-Term Memory; Bi-LSTM）と条件付き確率場を組み合わせたモデル Bi-LSTM-CRF を提案した．また Le ら [11] は，文字単位の情報を抽出するために畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）と Bi-LSTM-CRF を組み合わせた Bi-LSTM-CNN-CRF を提案した．

しかしこれらのモデルを学習するには，訓練データとしてフルアノテーションコーパスが必要であり，モデルが複雑になるほど必要な訓練データ数も増加する傾向にある．そのため，アノテーション作業を効率化することは重要であり，それに関連した研究についても紹介する．

一つには，能動学習が挙げられる．能動学習とは，人間とモデルとの対話的な学習方法であり，機械学習モデルがより多くの情報を得られるデータを選択し，そのデータのみを訓練データとして学習を行う手法である．これは，情報が少ないデータに対する無駄なアノテーション作業を減らすことができるため，アノテーション作業の効率化に繋がる [12], [13].

他には，クラウドソーシングという手法も挙げられる．クラウドソーシングとは，不特定多数のワーカに作業を委託してデータを集める手法である．アノテーションの品質が低くなる可能性があるものの，アノテーションに専門的知識を要さないタスクであればより多くの人間に作業を委託できるため，効率よくアノテーションデータを収集することができる [14].

しかし，これら二つの手法は，訓練データとして使用するデータ全てにアノテーション作業を行う必要があるため，一定

Spouse タスク	文章 S	They include Joan Ridsdale, a 62-year-old payroll administrator from County Durham who was hit with a €16,000 tax bill when her husband Gordon died.
	教示文 E	because the phrase “her husband” is within three words of person 2.
Disease タスク	文章 S	Young women on replacement estrogens for ovarian failure after cancer therapy may also have increased risk of endometrial carcinoma and should be examined periodically.
	教示文 E	because “risk of” comes before the disease.

図 2　BabbleLabel におけるタスクの例 [15]

のアノテーションコストを免れることはできない．一方，提案手法においては，自然言語教示を用いることで，一つの自然言語教示から獲得した情報から，複数のデータのアノテーションに活用することが出来る．

すなわち，提案手法においてはアノテーションコスト削減の問題解決に対して，使用する全訓練データに人手でアノテーションするという従来の手法とは異なった手法を，自然言語教示を用いることで可能にする．

2.2　自然言語教示による学習

本節では，自然言語教示を利用した研究について紹介する．Hancock ら [15] は，単純なラベル付与以上の情報を獲得するために各ラベル付与決定に対しての説明を自然言語文で表現し，それを分類器の学習のために使用したフレームワーク BabbleLabel を提案した．この研究では，二つの単語の関係性を問う二値分類問題を扱っている．例えば，アノテータに二人の人物が登場する文章を与え，その二人が婚約関係にあるのかを分類する Spouse タスクや化学物質名と病名が含まれる文章から，その化学物質が病気を引き起こす原因であるのかを分類する Disease タスクなどを扱った．図 2 には，この研究における各タスクと自然言語教示の例を示す．この研究では提案手法 BabbleLabel と，比較手法として一般的な二値分類モデルを使用し，それぞれの評価を F1 値で示した．BabbleLabel は訓練データに自然言語教示を用いており，二値分類モデルでは数値で表わされたものを訓練データとして実験が行われた．この実験で，BabbleLabel は自然言語文 30 件の訓練データから，数千件の訓練データを用いた二値分類モデルと同等の性能を示した (表 1）．

また，Srivastava ら [16] も自然言語教示を用いた分類モデルの性能を示した．この研究は，E メールを複数のタイプに分類し，その各タイプに見られる特徴を自然言語文によって表現することで分類モデルを学習した．評価方法は，未使用の E メールに対してどれほどの精度で分類できるかを F1 値で示し，ロジスティック回帰のような複数の既存の分類モデルと比較を行った．ロジスティック回帰に対しては，約 10 %の向上を示した．

これらの研究では，分類問題解決のための機械学習モデルを学習するために自然言語教示を利用しているが，本研究ではこの自然言語教示をフレーズ抽出タスク解決のために利用する．

3　手　　法

本章では，フレーズ抽出器の効率的な学習方法として自然言語教示を利用した手法について提案する．まず，3.1 節で提案手法の概要について説明する．3.2 節で提案手法である自然言語教示の利用について説明する．

表 1　BabbleLabel の実験結果 [15]

	BabbleLabel	二値分類モデル					
訓練データ数	30	30	60	100	300	1000	3000
F1 値	42.3	32.1	32.6	34.4	37.5	41.9	44.5

3.1　概　　要

1 章において，一般的なフレーズ抽出器の学習には多数の訓練データが必要であり，その訓練データの作成にはいくつかの問題点があることについて述べた．これらの問題点を解決するには，用意すべき訓練データの数を抑えることやアノテーションコストを軽減することが重要である．そのために，本研究では自然言語教示を利用する．提案手法は，大きく三つの部分に分けられ，順に詳細を述べる．

（1） 自然言語教示文を解析する CRF モデルの学習．
（2） Labeling Function（LF）から訓練データの生成．
（3） 点予測によるフレーズ抽出器の学習．

3.2　自然言語教示によるフレーズ抽出器の学習

本項では，提案手法について説明する．3.2.1 項では自然言語教示文を解析するための CRF モデルについて説明する．3.2.2 項では，CRF モデルによって解析された自然言語教示文からラベル付与のための関数（Labeling Function; LF）の作成について説明する．3.2.3 項では，Labeling Function によってラベル付与された訓練データから点予測によるフレーズ抽出器の学習について説明する．

3.2.1　自然言語文解析モデルの学習

自然言語教示文とは，アノテータがラベル付与の理由を自然言語文で表現したものである．図 3 は，アノテータが文書中のフレーズ Bryan Robson を人名と解釈した際の自然言語教示文であり，図 4 はフレーズ Grozny を人名ではないと解釈した際のものである．自然言語教示文を利用することで，単純なラベル付与以上の情報量を得られると考えられる．

始めに，自然言語教示文を解析するモデルを学習する必要があり，このモデルを学習するためにスロット・インテント推定方式を用いる．スロット・インテント推定方式とは，ユーザの発話理解を目的とし，対話システムモデルの学習などにも使用される手法である．対話システムモデルでは，与えられた自然言語文（ユーザ発話）に２つの処理を行う．１つは，ユーザ発

文章 S	Ravanelli aggravated ~~~ on Saturday by his manager Bryan Robson.
教示文 E	Becase "his manager" comes before "Bryan Robson"

図 3　人名であると解釈する自然言語教示文

文章 S	Alexander told ~~~ south of the Chechen capital Grozny, ~~~
教示文 E	Because the words "Chechen capital" come before "Grozny"

図 4　人名ではないと解釈する自然言語教示文

話からユーザの意図をを複数のカテゴリに分類する処理であり，この処理で得られる部分がインテントである．もう 1 つは，インテントに基づき，ユーザ発話の意味部分を表現している内容をスロットとして抽出する処理である．例えば,「東京から大阪までの距離を教えて」というユーザ発話からは，インテント（”距離”）とスロット（出発地=”東京”, 目的地＝”大阪”）が抽出され,「東京から北海道までの交通費を教えて」というユーザ発話からは，インテント（”費用”）とスロット（出発地＝”東京”, 目的地＝”北海道”）が抽出されるということである．これを本研究に適用させると，対話システムモデルにおいてのユーザ発話が自然言語教示文である．

このタスクはインテントを推定する分類タスクと，スロットを推定する系列ラベリングタスクの組み合わせとして扱うことができることから，機械学習モデルの学習が必要である．インテント推定のための訓練データには，自然言語教示文を用い，モデルはロジスティック回帰を採用する．また，スロット推定のための訓練データには IOB タグを付与したものを用い，モデルは条件付き確率場（Conditional Random Field; CRF）を採用する．そして，インテントごとの CRF モデルを学習し，フレーズ抽出したいドメインに関する自然言語教示文を入力とするとスロットが出力として得られ，この出力されたスロットをこの文の解釈とみなす．

次に，スロット・インテント推定方式を用いた具体的な流れを説明する．まず，ロジスティック回帰モデルを用いて，各文をインテントごとに分類する．提案手法において，インテントは以下の 4 つに分類され，各文はこれら 4 つのどれかに分類される.「comesbefore」は，抽出するフレーズの前に特定の単語が来ることを表し,「comesafter」は，抽出するフレーズの後に特定の単語が続くことを表す．また.「isA」は，その単語の意味を直接表し,「other」は，提案手法において解釈しない，解釈出来ないことを表す．インテントが「comesbefore」,「comesafter」,「isA」に分類された文は，それぞれのインテントごとの学習済み CRF モデルで予測をすることで，スロット推定を行う．スロット推定の結果は図 5 に示しており，スロットを表す部分の開始位置に B タグ，途中位置に I タグ，スロットを表していない部分に O タグが付与される．

スロット推定により抽出される部分文字列は，インテントごとに異なり，以下のことを示している．また，インテント「other」については，提案手法において解釈しない，解釈出来ない文であるためスロット推定も行わない（表 2）.

- インテント「comesbefore」のスロット推定された文字列は，文章中において抽出したいフレーズの前の文字列を示す.
- インテント「comesafter」のスロット推定された文字列は，文章中において抽出したいフレーズの後の文字列を示す.
- インテント「isA」のスロット推定された文字列は，文章中において抽出したいフレーズそのものを表す文字列を示している.

例えば,「My husband John like playing the guitar .」という文章中の「John」という単語が人名であるかどうかをアノテータに問うと，アノテータは「John」を人名と解釈し，その理由を,「Because my husband comes befre John」という自然言語教示文で表現したとする．この文は,「John」の前に位置する「My husband」が「John」を人名と解釈できる特定のフレーズであることを意味しているため，学習済みモデルを使用して予測をすることで，インテントが「comesbefore」，スロットは「my husband」が抽出され，この自然言語教示文の解釈とする.

文字列 x	Because	my	husband	comes	before	John	.
ラベル列 y	O	B	I	O	O	O	O

図 5　提案手法におけるスロット推定

3.2.2 Labeling Function

本項では，Labeling Function（LF）について説明する．LF とは，自然言語教示から得られた情報を用いて作成する関数であり，ラベル付与のために使用する．提案手法においては，少量の自然言語教示文から，LF を作成し，それらを未使用コーパスに適用することで，機械的にラベル付与されたコーパスを大量に生成する.

LF の生成は，インテントが「comesbefore」または「comesafter」であるか「isA」であるかによって異なる処理を行う.

インテントが「comesbefore」または「comesafter」に分類された自然言語教示文のスロット推定結果は，それぞれ抽出したいフレーズの前後に特定の文字列が続くことを意味しているので，そのスロットを引数とした関数を生成し，生コーパスに適用させることで，生コーパス中にそのスロットと同様の文字列が出現した際の前後 (その関数が comesbefore から生成されているのであれば，前．comesafter から生成されているのであれば後ろ) のフレーズを抽出することができる.

また,「comesbefore」に分類された自然言語教示文と「comesafter」に分類された自然言語教示文は，文の構造が似ていることからそれぞれ逆のインテントに分類されてしまっているケースが考えられる（本来は,「comesbefore」であるべきだが,「comesafter」に分類されているような場合）．これを解決するために，インテント「comesbefore」と分類されているものでも同様のスロットを用いて，インテント「comesafter」である場合の LF も作成する (逆も同様である)．これは，インテント分類の結果を無視しているように思えるが，LF 生成後にフィルターを通すことで不適切な LF を取り除けるため問題ない．フィルターについては，この項の後半で説明する.

次にこれら二つのインテントとは異なった処理を行う，インテント「isA」について説明する．例えば,「EU rejects German call to boycott British lamb .」という文章中の「EU」という単語が人名であるかどうかをアノテータに問うと，アノテータは「EU」を人名ではないと解釈し，その理由を「It is the name of an organization」という自然言語教示文で表現したとする．これは，アノテータが「EU」という文字列の意味を知っており，この文字列は組織名を意味しているため人名ではないと解釈したと考えることができるため，インテントを「isA」に分類し，スロットは「organization」が抽出される．そしてこの自然言語教示文から.「organization」であるものは人名ではないと判断す

表 2　各インテントとスロット

インテント	スロット	インテント	スロット
comesbefore	抽出対象のフレーズの前の文字列	isA	抽出対象のフレーズそのものを表す文字列
comesafter	抽出対象のフレーズの後の文字列	other	ー

ることができるため，辞書のような知識源を有しており使用できる環境においては,「EU」の他に,「organization」であると判断できるフレーズに関しては「EU」同様に人名ではないと解釈することができる．

こうして機械的に多数の LF が作成されると，この中には適切な LF と不適切な LF が存在することになる．そこで，不適切または不要な LF を出来るだけ除外するために二つのフィルターに通す．それぞれは，意味的フィルター（Semantic Filter）と重複フィルター（Duplication Filter）と呼ぶことにする．

意味的フィルター（Semantic Filete）は，その LF が元の自然言語教示文を正しく解釈できているかどうかを識別するものであり，正しく解釈されていない LF は除外される．具体的には，作成された LF を元の自然言語教示文に適用させ，正しくラベル付与されているかどうかを確認する．

重複フィルター（Duplication Fileter）は，関数的に同等の役割を果たす LF を除外するものである．

インテント「comesbefore」とインテント「comesafter」から獲得した LF と，インテント「isA」から抽出されたフレーズを用いて知識源から獲得した知識の二つを使用して未使用コーパスにラベル付与を行う．

ここまでの流れを図 6 を用いて説明する．まず，生コーパス「My husband James won a golf championship in America」に対して，獲得した LF を全て適用させる．すると，獲得した全ての LF の中で図中の三つの LF が当てはまったとする．一つ目の LF から「My husband」に続く「James」が人名であること，二つ目の LF から「championship」の前の「golf」が人名でないこと，三つ目の LF から知識源を参照し,「Japan」と同様な国名を表す表現は人名でない（つまり,「America」も人名でない）ことが分かり，それぞれに適切なラベル付与を行うことができる．その結果，生コーパスを LF によって一部ラベル付与された部分的アノテーションコーパスへと変換することができ，同様の処理を全コーパスに用いることで，訓練データを作成する．部分的アノテーションとは，何かしらのラベルが付与されている単語とラベルが付与されていない単語が共に存在するアノテーションのことである．

3.2.3　点予測によるフレーズ抽出器の学習

3.2.2 項より部分的アノテーションがなされた訓練データを利用して，点予測によるフレーズ抽出器の学習を行う．点予測とは，Neubig ら [17] が提案したもので，文字列 $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_m \rangle$ から単語列 $\boldsymbol{w} = \langle w_1, w_2, \ldots, w_n \rangle$ を予測する単語分割器の学習に使用された．従来の単語分割器の学習にはフルアノテーションコーパスを必要とする隠れマルコフモデル（HMM）や条件付き確率場（CRF）が一般的な手法とされていたが，この点予測は部分的アノテーションコーパスでの学習を可能とした．その後小林ら [18] が，固有表現抽出器の学習に活用出来るように拡張を行った．

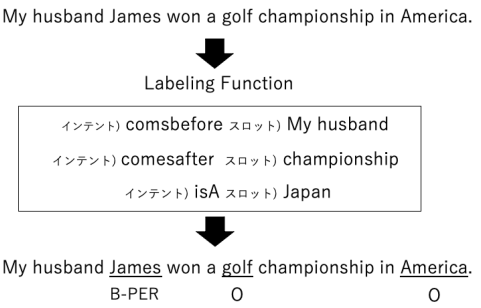

図 6　LF を用いての部分的アノテーションコーパス作成

点予測では，部分的アノテーションコーパスを使用するため条件付き確立場（CRF）のような系列全体を考慮するような予測とは異なり，ある単語 x_i に対してラベル y_i を予測するので，多クラス分類問題と考えられる．多クラス分類問題を解決する機械学習手法はいくつかあるが，小林らは多クラスロジスティック回帰を使用している．

ロジスティック回帰とは，二値分類を扱う線形分類器の一つであり，入力 x が二つのクラス $y \in \{+1, -1\}$ に属する確率を求めるもので，これを式（1）に示す．$\boldsymbol{f}(x)$ は素性関数 $f(x)$ からなる素性ベクトルであり，$\boldsymbol{w}$ は素性ベクトルに対しての重みベクトルである．

$$P(y|x) = \frac{1}{1 + \exp\left(-y\boldsymbol{w}^T \boldsymbol{f}(x)\right)} \tag{1}$$

訓練データ $(\boldsymbol{x}, \boldsymbol{y}) = (\langle x_1, x_2, \ldots, x_n \rangle, \langle y_1, y_2, \ldots, y_n \rangle)$ に対して，尤度 L を最大化させることで重みベクトル $\boldsymbol{w}$ の最適化を行い，尤度 L は以下の式で定義される．

$$L(\boldsymbol{w}) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}; \boldsymbol{w}) \tag{2}$$

そして，この尤度 L の対数を取ったものを対数尤度 l と呼び，最大化を行う．式は以下で定義される．この対数尤度は最適化手法で良く知られている勾配降下法を用いることで，重みベクトル $\boldsymbol{w}$ を計算することができる．

$$l(\boldsymbol{w}) = \log \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}; \boldsymbol{w}) = \sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)}; \boldsymbol{w}) \tag{3}$$

多クラスロジスティック回帰においては，入力 x を L 個のラベルの中から一つのラベルに推定するモデルであり，各ラベル y に対してソフトマックス関数を適用させ，入力 x が各ラベルに属する確率を求める．

$$P(y|x) = \frac{\exp\left(\boldsymbol{w}_y^T \boldsymbol{f}(x)\right)}{\sum_{y' \in y} \exp\left(\boldsymbol{w}_{y'}^T \boldsymbol{f}(x)\right)} \tag{4}$$

そして小林らの点予測では，ある単語 x_i に対して m を窓幅とすると，x_i の周辺単語前後の周辺単語である $x_{i-m+1}, \ldots, x_{i-1}$ と $x_{i+1}, \ldots, x_{i+m}$ を入力としてラベル y_i を予測しており，以下の

表 3　CoNLL データセット

名称	文章数
データ A	1000
データ B	11636
データ C	1405

表 4　自然言語教示文

名称	文章数
教示文 a	1000
教示文 b	1000
教示文 c	1000

式で定義される．

$$P(y|x_{i-m+1},\dots,x_i,\dots,x_{i+m}) = \frac{\exp\left(\boldsymbol{w}_y^T \boldsymbol{f}(x_{i-m+1},\dots,x_i,\dots,x_{i+m})\right)}{\sum_{y'\in\mathcal{Y}} \exp\left(\boldsymbol{m}_{y'}^T \boldsymbol{f}(x_{i-m+1},\dots,x_i,\dots,x_{i+m})\right)} \quad (5)$$

提案手法においても，部分的アノテーションコーパスを使用したフレーズ抽出を行うことから，この点予測を用いることにする．また，その他素性関数の設計なども小林らの既存手法に従い実験を行う．

4　実験と評価

本章では，提案手法の実験と評価について述べる．まず，4.1 節で評価実験に用いるデータセットについて説明する．次に，4.2 節で実験方法について述べ，4.3 節で実験結果・考察を示す．

4.1　実験用データ

評価実験に用いるデータセットは，CoNLL-2003 English dataset [6] を用いる．このデータセットをトレーニングデータセット，生データセット，テストデータセットの三つに分割し，それぞれをデータ A, データ B, データ C とする．文章数は表 3 で示す．データ A は，三人のアノテータに理由を聞き，自然言語教示文を獲得するために使用した．それぞれのアノテータから獲得した自然言語教示文は，それぞれ教示文 a, 教示文 b, 教示文 c とし，データ数は表 4 で示す．教示文 a，教示文 b はの自然言語文解析モデルの学習に使用され，このモデルによって教示文 c を解析することにより，Labeling Function（LF）を獲得する．データ B は，教示文 c から作成された LF によってラベル付与される，未使用の生データセットとして使用した．データ C は，提案手法の評価実験に使用され，モデルの性能を計測した．

4.2　実 験 方 法

本実験では，抽出すべきフレーズを人名とし，自然言語教示を利用し，効率的な学習器の学習を行い評価する．モデルの比較のために，三つのフレーズ抽出器を学習させる．

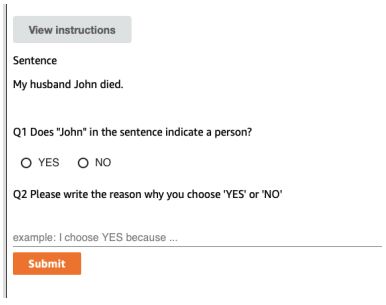

図 7　AmazonMechanicalTurk の使用例

まず提案手法では，トレーニングデータ A を使用して三人のアノテータから特定のフレーズが人名であるかどうかの回答と自然言語教示文を獲得し，それぞれをデータ a, データ b, データ c とする．本実験では，クラウドソーシングである Amazon Mechanical Turk を利用した（表 7）．データ a,b を利用してロジスティック回帰モデル L と CRF モデル C の学習を行う．この学習の流れは 3.2.1 項に従う．次に，データ c に対しモデル L を使用することで各文のインテントを予測し，そのインテントごとにモデル C を使用してスロット推定を行う．これらのインテントの予測，スロット推定結果より 3.2.2 項に従い Labeling Function（LF）を作成する．インテントが「comesafter」,「comesbefore」に分類された自然言語教示文はスロットを元に関数化する．インテントが「isA」に分類された自然言語教示文に関しては，知識源を有している時にそれを参照する．今回は，この知識源として WordNet を利用する．WordNet とは英語の概念辞書であり，英単語は synset と呼ばれる概念と結びついており参照可能である．WordNet の仕様についてはオープンソースソフトウェアとして公開されている．

そしてこれらの LF と WordNet から獲得した知識を使用して，未使用コーパスデータ B にラベル付与を行い，未使用コーパスデータ B をラベル付与済みコーパスデータ B’ に変換する．最後にこれらデータ B’ を訓練データとして，点予測によるフレーズ抽出器モデル M の学習を行う．評価方法は，データ C をテストデータとしてモデル M で予測した結果と，テストデータの正解を照らし合わせ，各文字列ごとの予測結果を Precision, Recall，F1 値で示す．

次にベースライン手法として，アノテータから特定のフレーズが人名であるかどうかの回答のみを獲得し，自然言語教示文を利用せずにフレーズ抽出器の学習を行う．ベースライン手法ではまず，トレーニングデータ A を使用し，三人のアノテータから特定のフレーズが人名であるかどうかの回答を獲得する．その獲得した回答から，その回答と対応しているトレーニングデータにそのフレーズが人名であるかどうかのラベル付与を行い，ラベル付与されたデータをデータ A’ とする．データ A’ を用いて，点予測によるフレーズ抽出器モデル M の学習を行う．評価方法は，提案手法と同様にデータ C をテストデータとして予測を行い．Precision，Recall，F1 値で示す．

既存手法としては，固有表現抽出（Named Entity Recognition; NER）手法で用いられるフルアノテーションコーパスを使用したモデルを学習する手法を用いる，特に今回の実験では Bi-LSTM-CRF を用いる．Bi-LSTM-CRF とは，再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）の一種である Bi-LSTM と CRF を組み合わせたものである．RNN には，系列の要素が順番に渡されるので，ある時点における情報を記憶し，その記憶した情報をその後の学習に使用することができる．ただ，長期間における情報の記憶が困難であったため，それを可能にしたのが LSTM（Long Short Term Memory）である．また Bi-LSTM とは系列データを双方向からの LSTM を行ったものであり，Bi-LSTM-CRF とは，Bi-LSTM の結果を系列全体を考

表 5　獲得した自然言語教示文

インテント	自然言語教示文
comesbefore	Becase ”his manager” comes before ”Bryan Robson”
comeafter	Because ’said’ comes after Colin Hope.
isA	it’s a soccer team
other	Two names in sentence both are players

慮して扱うものである [19].

今回の実験では，データ B を用いてこの Bi-LSTM-CRF の学習を行い，評価方法は，提案手法と同様にデータ C をテストデータとして予測を行い，Precision，Recall，F1 値で示す.

4.3　実験結果・考察

4.2 節の実験方法に従って実験を行う．本実験では，アノテータに特定のフレーズが人名であるかどうかを問いその理由を自然言語教示という形で利用した提案手法によるフレーズ抽出器モデルと，ベースライン手法としてアノテータに特定のフレーズが人名であるかどうかを問うが自然言語教示を利用しないフレーズ抽出器モデルと，フルアノテーションコーパスを用いた NER 手法によるモデルの三つの比較を行った.

それぞれの結果は，表 7，表 8，表 9 に示しており，これらは各文字列を人名である・ないと予測したモデルの Precision，Recall，F1 値による評価である．また，表 5 には獲得した自然言語文を示す.

表 7，表 8，表 9 は全て同じ項目で構成されており，縦軸は各評価項目であり，横軸はタスク回答数である．表 7 におけるタスク回答数とは，特定のフレーズが人名であるかどうかの回答と自然言語教示を合わせたものを一つのタスクと数えている．表 8 におけるタスク回答数とは，自然言語教示を使用していないため，特定のフレーズが人名であるかどうかの回答のみを一つのタスクと数えている．表 9 におけるタスク回答数とは，モデルの学習に使用したフルアノテーションコーパスの数である.

まず表 7 と表 8 より，提案手法とベースライン手法の F1 値をタスク回答数ごとに比較してみるとおおよその場合で，提案手法の値がベースラインの手法を上回ることが示された．次に表 7 と表 9 を比較すると，提案手法のタスク回答数 1000 件において NER 手法と同等の性能を示した．提案手法と NER 手法のタスク回答数は同じであるが，提案手法のタスクは一つのフレーズを人名であるかどうかを回答しそれを自然言語教示で表現する一方，NER 手法は一文をフルアノテーションする必要があるため，学習コスト面では異なるとも考えられる.

これらの実験結果が示されたのには，二つの要因が考えられる．一つは，提案手法では LF や WordNet を利用することで訓練データ数を増加させたことにあると考えられる．ベースライン手法・NER 手法では，タスク回答数がそのままモデルの訓練データ数となっているのに対し，提案手法では，どのタスク回答数に対しても生データセットであるデータ B（11636 件）を使用するので，生データセットの数が確保されていれば，少量のタスク数からでも十分な訓練データを用意することができる．しかし，タスク回答数 20，50，100 のような少量の自然言語教示文からは LF を獲得することはできず，WordNet からも数十語しか獲得できなかったため，あまり効果的な学習は行えなかった.

二つ目は，自然言語教示文から作成された LF や WordNet から獲得した語彙により，訓練データに追加でラベル付与を行えたことにあると考えられる．まず，本実験で獲得したインテントと LF を表 6 に示す．見方としては，LF の一項目「インテント」は，その関数のインテントを値で表現しており，0 は「comesafter」，1 は「comesbefore」，2 は「isA」であることを示している．二項目「タイプ」はその関数のタイプを表現しており，”B”が人名タグを付与するための関数，”O”が人名ではない，つまり O タグを付与するための関数であることを示している．三項目「スロット」はスロット推定により抽出された文字列を表現しており,「comesafter」,「comesbefore」であればフレーズを抽出するための特定の文字列を示し,「isA」であれば，WordNet を参照する際に用いる文字列であることを示している．そしてこれらは，事前に用意した関数の引数として実際に使用される.

インテントごとの説明をすると，インテント「isA」からは，人名でないフレーズを獲得するための情報を得ることができ，WordNet を参照して語彙を増やした．例えば,「organization」だから人名でないと判断される時には，WordNet に「organization」を問い合わせその下位概念の単語まで探索することで,「EU」や「WHO」のようなフレーズは人名ではないという知識を獲得することができる．このように WordNet を利用することで，インテント「isA」に分類された文を扱うことができた．またインテント「comesafter」とインテント「comebefore」に分類された文は，LF へと変換されたが，今回の人名抽出というタスクにおいて効果的に働いたものは少なかった．提案手法のタスク回答数 1000 件においては，獲得した自然言語教示文から機械的に LF を作成し 237 の LF を獲得した．その後フィルターにかけ残った LF は 18 であった．この LF に関しては，いくつか考えられることがある.

一つは，今回実験を行った人名抽出タスクにおいてはこのような結果であったが，抽出対象を変更すると LF にも変化が見られるかもしれないということである．インテント「comesafter」とインテント「comebefore」は抽出対象のフレーズの前後に，何か特定のフレーズが置かれているのではないかという考えからくるものであるため，抽出対象によって変化が見られる可能性も考えられる．そのために，異なる抽出対象を選択したタスクを提案手法と同様のフレームワークで行い，これらの LF の有効性を検証する必要がある.

表 6　自然言語教示から変換した LF

	LF		
インテント	インテント	タイプ	スロット
comesafter	0	”O”	”championship”
comesbefore	1	”B”	”president”
isA	2	”O”	”city”

表 7　提案手法の実験結果

タスク回答数	20	50	100	200	400	1000
Precision	1.00	1.00	0.98	0.64	0.58	0.64
Recall	0.18	0.13	0.22	0.55	0.77	0.83
F1 値	0.30	0.24	0.36	0.59	0.66	0.72

表 8　ベースライン手法の実験結果

タスク回答数	20	50	100	200	400	1000
Precision	0.07	0.31	0.32	0.36	0.37	0.44
Recall	0.34	0.52	0.67	0.69	0.82	0.80
F1 値	0.11	0.39	0.43	0.48	0.51	0.57

表 9　フルアノテーションコーパスを用いた NER 手法の実験結果

タスク回答数	20	50	100	200	400	1000
Precision	0.16	0.29	0.24	0.58	0.66	0.83
Recall	0.01	0.04	0.20	0.34	0.52	0.57
F1 値	0.01	0.06	0.22	0.43	0.58	0.68

またもう一つは，ラベル付与のための関数にしないということである．今回の LF のように関数化してしまうと，正解か不正解かが明確なため，ラベルを付与するかしないかの二択になってしまう．そこで，これらのインテントから獲得した情報を関数化するのではなく，フレーズ抽出器自体の特徴量に設計するという手法が考えられる．この手法は，フレーズ抽出器を学習していく上で，特定のフレーズを抽出するのに関わりの強い特徴量であればその特徴量に対する重みが増やされ，あまり関わりのない特徴量であればその特徴量に対する重みが減らされるだろう．

5 おわりに

本研究では，訓練データの削減やアノテーションコストを軽減するために自然言語教示を用いたフレーズ抽出器の学習手法について提案した．自然言語教示を用いることで，少量の訓練データから大量の擬似訓練データを作成し，アノテータが用意すべき訓練データ量を減らすことができる可能性を示した．その中で効果を発揮した部分とそうでない部分も明らかになり，今後の展開について考えることにも繋がった．具体的には，本研究での提案手法と同様の手法を用い，実験で行った人名抽出タスク以外のタスクに取り組む必要がある．始めの段階では，固有表現という観点から組織名抽出タスク，地名抽出タスクなどが挙げられ，その後は病名や化学物質名のような専門用語抽出タスクへと応用するべきである．また，提案手法で提案した LF が他のカテゴリの抽出タスクにおいてもあまり効果が見られないのであれば，フレーズ抽出器の特徴量として扱った場合のモデルはどういった性能を示すのかを検討する必要があると考えられる．

謝　辞

本研究の一部は，JSPS 科研費（課題番号 19K20333）および JST CREST (#JPMJCR16E3) AIP チャレンジの助成によって行われた．

文　献

[1] S. Morwal, N. Jahan, and D. and Chopra. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC)*, Vol. 1, No. 4, pp. 15–23, 2012.

[2] A. Ekbal and S. Bandyopadhyay. A hidden markov model based named entity recognition system: Bengali and hindi as case studies,. *Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*, Vol. 4815, pp. 545–552, 2007.

[3] 土田正明, 松井藤五郎, 大和田勇人. 論文からの専門用語抽出とウェブを用いた用語説明生成. 日本ソフトウェア科学会大会講演論文集 日本ソフトウェア科学会第 21 回大会, pp. 77–77. 日本ソフトウェア科学会, 2004.

[4] Erik F Sang and Jorn Veenstra. Representing text chunks. *arXiv preprint cs/9907006*, 1999.

[5] 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.

[6] F. Erik, Tjong kim sang, and Fien De Meulder. Introduction to the conll-2003 shared task:language-independent named entity recognition. *InProceedings of the Seventh Conference onNatural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

[7] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[8] Arjun Das and Utpal Garain. Crf-based named entity recognition@ icon 2013. *arXiv preprint arXiv:1409.8008*, 2014.

[9] K. Riaz. Rule-based named entity recognition in urdu. *Association for Computational Linguistics 2010, Uppsala, Sweden*, pp. 126–135, 2010.

[10] R. Alfred, L. C. Leong, and et al. A rule-based named-entity recognition for malay articles. *Lecture Notes in Computer Science, Berlin, Heidelberg*, Vol. 8346, pp. 288–299, 2013.

[11] Anh Le and Mikhail Burtsev. A deep neural network model for the task of named entity recognition. *International Journal of Machine Learning and Computing*, 2018.

[12] Burr Settles, Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP ’08, pp. 1070–1079, Stroudsburg, PA, USA, 2008. Asso- ciation for Computational Linguistics.

[13] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. CoRR, Vol.abs/1707.05928, 2017.

[14] 森嶋厚行. クラウドソーシングが不可能を可能にする—小さな力を集めて大きな力に変える科学と方法—. 2020.

[15] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1884–1895, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[16] Shashank Srivastava, Igor Labutov, and Tom Mitchell. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1527–1536, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[17] Graham Neubig. 点推定と能動学習を用いた自動単語分割器の分野適応. 言語処理学会年次大会, 2010, 2010.

[18] 小林滉河, 若林啓. 点予測と能動学習を用いた固有表現抽出の提案. *https://db-event.jpn.org/deim2019/post/papers/339.pdf*, 2019.

[19] 手塚太郎. しくみがわかる深層学習. 朝倉書店, 6 2018.