B0TB2127

卒業論文

Wikipedia 記事への拡張固有表現ラベルの多重付与

鈴木正敏

2016 年 3 月 31 日

東北大学
工学部 情報知能システム総合学科

Wikipedia 記事への拡張固有表現ラベルの多重付与*

鈴木正敏

内容梗概

本論文では、Wikipedia の個々の記事に対して細粒度の固有表現分類のラベルを付与するタスクに取り組む。分類を細かくした際に生じるデータスパースネスの問題に対処するため、本研究ではニューラルネットを用いたマルチタスク学習によって、全てのクラスのラベル付与を同時に学習することを提案した。また、分類器の訓練に用いる素性空間が疎になることに対処するため、Wikipedia 本文全文から記事のリンクの分散表現を Skip-gram モデルで学習し、分類器の訓練に用いた。実験の結果、提案した手法により、既存研究を再現したベースラインと比較して、事例ベースの F 値でおよそ 5 ポイントの改善が見られた。特に、比較的記事数の少ないクラスにおいて、分類性能の大きな向上が見られた。

キーワード

固有表現分類、Wikipedia、マルチタスク学習

*東北大学 工学部 情報知能システム総合学科 卒業論文, B0TB2127, 2016 年 3 月 31 日.

目 次

図 目 次

表 目 次

1　はじめに

本研究では、Wikipedia の記事に対して固有表現クラスのラベルを自動的に付与するタスクに取り組む。

人や物、出来事に関する知識は、固有名詞や時間表現、数値表現といった固有表現によって表される。大規模なオンライン百科事典である Wikipedia は、このような固有表現に関する知識源として、その価値が注目されている。一方で、その記事は自然言語で書かれているため、必ずしも計算機で扱いやすいような形式にはなっておらず、構造化が必要である。

知識の構造化においては、個々の事物（エンティティ）に対して「人名」「地名」などといった固有表現分類に関する知識を構築することが重要になる。固有表現分類は、似た意味的役割を持つ固有表現をグループ化したクラスであり、このクラスに基づいてエンティティが持つ属性やそれらの間に定義されうる関係を整理した知識ベースは、ファクトイド型質問応答や知識ベースに基づく推論のための基盤知識として重要である。

Wikipedia の記事に対して固有表現分類を付与する既存研究はいくつか存在する (Chang et al., 2009; Dakka and Cucerzan, 2008; Higashinaka et al., 2012; Tardif et al., 2009; Toral and Muñoz, 2006; Watanabe et al., 2007) が、そのほとんどは、3 から 15 クラス程度の比較的粗い分類体系に基づくものである。その一方で、より細かい粒度での分類は、エンティティリンキング (Ling et al., 2015) や質問応答 (Mann, 2002) といった種々の自然言語処理のタスクにおいて有用であることが知られている。そこで本研究では、細かい粒度の分類体系に基づいた、Wikipedia 記事の分類に取り組む。

既存研究の多くは機械学習に基づく手法で記事の分類を行っているが、それを細かい粒度での分類にそのまま適用しようとすると、データスパースネスの問題が生じる。例えば、「日本」「富士山」「東京ドーム」といった記事は、従来の粗い粒度の分類では全て「地名」という分類ラベルを付与していたが、細かい粒度の分類では、それらの記事に対して「国名」「山地名」「競技施設名」といった分類ラベルをそれぞれ付与することになる。分類の粒度を細かくしたことで、1 クラスあたりの事例数が少なくなり、クラスごとに十分な数の訓練データを用意する

ことが難しくなる。この問題に対処するため、本研究では2つの手法を提案する。

1つは、隠れ層を持つニューラルネットを用いて、全てのクラスのラベル付与を同時に学習することである。このモデルでは、学習される隠れ層のパラメタが全てのクラスで共有されることになる。これにより、分類先のクラス間の依存関係が学習され、分類性能の向上につながることが期待される。

もう1つの提案手法は、Wikipedia 内のリンクの周辺文脈を素性として用いたことである。既存研究では、記事名の bag-of-words といった離散的な素性が分類器の訓練に用いられることが多かったが、素性空間が疎になりやすいという問題があった。分類に有効な素性を追加するため、我々は Wikipedia では記事同士が相互にリンクされていることに着目し、リンクの周辺文脈がリンク先の記事の分類に有効なのではないかと考えた。リンク元の周辺文脈を分類に用いるという手法自体はすでに存在するものの (Dakka and Cucerzan, 2008) 、既存研究では、周辺文脈を bag-of-words で表現していたため、素性空間の次元数が非常に大きくなり、結果として分類精度の向上に繋がらなかったことが報告されている。これに対して、我々は Skip-gram モデル (Mikolov et al., 2013b) に基づく手法により、Wikipedia 本文全文から、記事のリンクの分散表現を獲得し、低次元かつ値が連続的な素性として分類に用いた。

以上2つの提案手法を用いて、日本語 Wikipedia の記事に対して、200 クラスからなる「関根の拡張固有表現階層」(Sekine et al., 2002) のラベルを付与する実験を行った。その結果、既存研究を再現したベースライン手法と比較して、事例ベースの F 値が 4.97 ポイント向上した。また、特に比較的事例数の少ないクラスにおいて分類性能が大きく向上した。

本研究のタスクの概観図を図1に示す。

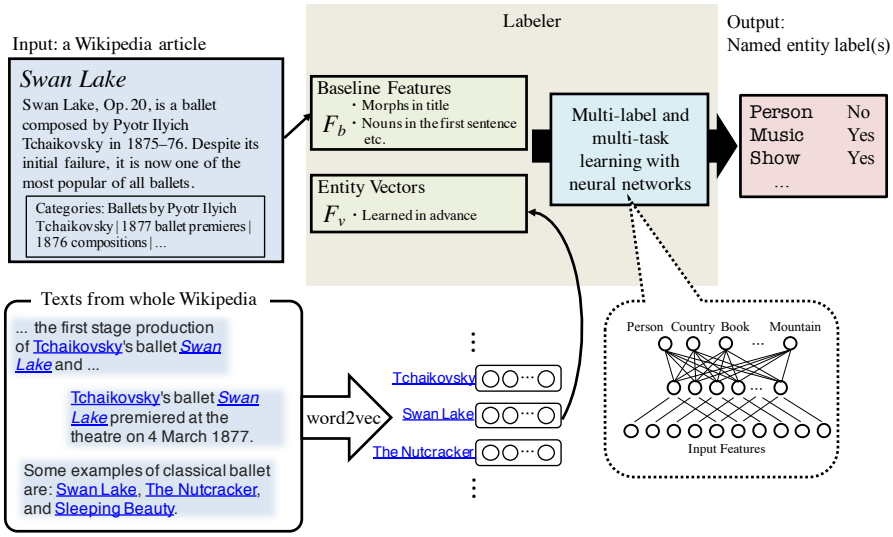


図 1: Wikipedia 記事に対する拡張固有表現ラベルの多重付与

2 関連研究

Wikipedia の記事に固有表現のラベルを付与するタスクは、Wikipedia を基に固有表現の分類付き辞書を作るタスクと共通する部分が大きい。ここでは、それらタスクに対する既存の取り組みについて述べる。

(Toral and Muñoz, 2006) は、`Location`、`Organization`、`Person` という 3 つのクラスについて、記事の本文に含まれる名詞がどのクラスに関連するかを WordNet を用いて分類し、クラスごとにそれらの名詞の数を数えることで記事のクラスを決定する手法を提案した。(Dakka and Cucerzan, 2008) は、ACE (Doddington et al., 2004) で用いられていた `PER`, `ORG`, `LOC`, `MISC` の 4 クラスを対象に、記事本文や表に含まれる語の bag-of-words と記事のリンク元の周辺単語の bag-of-words を素性に用いて、ナイーブベイズおよび SVM による教師あり学習による分類を行った。(Watanabe et al., 2007) は、Wikipedia の記事ページの HTML 構造から、アンカーテキストの出現の依存関係を反映したグラフ構造を作り、条件付き確率場というグラフベースの手法により、グラフ構造中のアンカーテキストでリンクされている記事を関根の拡張固有表現階層のうちの 13 クラスに分類した。その他にも、(Tardif et al., 2009) や (Chang et al., 2009) のような取り組みがある

が、いずれも、分類先のクラスが数〜十数クラスの粗いものである。

一方で、(Higashinaka et al., 2012) は、教師あり学習に基づいて、Wikipedia の記事を関根の拡張固有表現階層の約 200 クラスに分類することを試みた。彼らは、記事のタイトルや本文、カテゴリ情報や Infobox のテンプレートなどから分類に有効な素性を検討、抽出し、クラスの数だけロジスティック回帰による 2 値分類器を学習して、分類器の出力確率が最も大きいクラスを分類結果とする、という手法をとった。

さらに近年は、YAGO (Suchanek et al., 2007) や DBpedia (Auer et al., 2007) といった、Wikipedia の記事に対して、単純なヒューリスティクスや人手で整備されたルールに従ってラベルを付与する取り組みも存在する。しかしそれらの手法は、記事に付与されたメタデータに強く依存しており、ルールのカバレッジやメタデータの不足に対して問題がある (Aprosio et al., 2013)。

3　固有表現階層

本研究では、固有表現のオントロジとして、「関根の拡張固有表現階層」(Sekine et al., 2002) を用いた。これは、特定のドメインに依存しない固有表現の分類として 200 のクラスを定義したものであり、それぞれのクラスは 3 レベルの階層構造の中に位置している。そして、そのほとんどのクラスについては、そのクラス固有の属性が定義されている。例えば、「山地名」というクラスに対しては、「標高」や「登頂者」といった属性が定義されている。

本研究で関根の拡張固有表現階層を用いた理由は、クラスや階層構造の定義が、少数の人によって集中的にコントロールされているからである。Wikipedia 内で用いられているカテゴリや、DBpedia で定義されているオントロジは、不特定多数の人からなるコミュニティによって管理されているものであるが、分類の粒度やカバレッジに関して、必ずしも適切であるとは言えない。例えば、DBpedia では `AmericanFootballLeague` や `NarutoCharacter` といった過度に具体的なクラスが存在する一方で、`Medicine` のような、それよりも下位のクラスが存在しないような、範疇の広いクラスも存在する。Wikipedia のカテゴリについて言えば、ある記事にどのカテゴリを付与するかは記事の執筆者次第であり、カテゴリ付与の一貫性やカバレッジが保障されているとはいえないものになっている。

ところで、Wikipedia 記事の分類というタスクの実際を考えると、通常の多クラス分類問題のように、全てのクラスの中から最も適切な分類を 1 つだけ選ぶ、という設定は必ずしも適切であるとは言えない場合がある。例として、次の記事を考える。

> **記事名**: 世界の中心で、愛をさけぶ
>
> **記事本文**: 『世界の中心で、愛をさけぶ』(せかいのちゅうしんで、あいをさけぶ) は、日本の小説家・片山恭一の青春恋愛小説である。小学館より 2001 年 4 月に刊行。通称「セカチュー」。2004 年以降、漫画化、映画化、テレビドラマ化、ラジオドラマ化、舞台化されている。…

この記事に対しては、「文学名」「番組名」「映画名」といった複数のラベルを付与するのが妥当である。他にも、「ウルトラマン」(「番組名」と「キャラクター名」) や「トウモロコシ」(「植物名」と「食べ物名_その他」) など、記事が複数

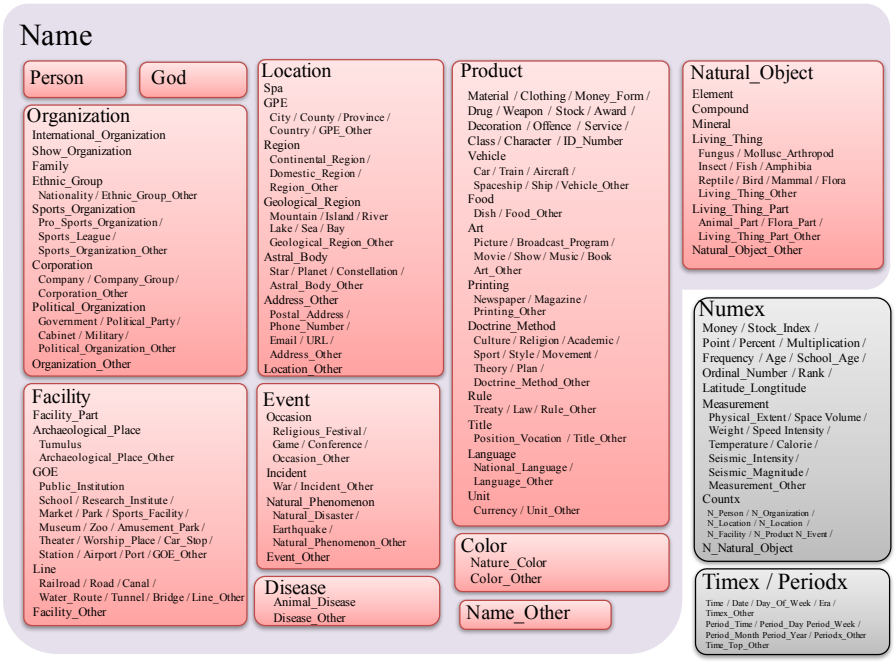


図 2: 関根の拡張固有表現階層で定義されているクラス

のカテゴリに属するケースは少なくない。実際、実験の章で述べる正解ラベルの統計を調べると、約 4.6 %の記事に複数のラベルが付与されていた。

本研究では、このような記事に対して妥当なラベル付与を行うため、各記事に対して複数のラベル付与を認めるマルチラベル分類としてタスクに取り組んだ。

4　モデル

最も単純なマルチラベル分類の実現方法の 1 つに、クラスの数だけ、そのクラスのラベルを付与するかどうかを判別する 2 値分類器を作り、それらを文書に対して適用した結果、出力が正となったすべてのクラスのラベルをその文書に付与するという手法がある（図 3a)。この手法では、あるクラスのラベルを付与するために学習される情報が他のクラスのラベル付与に影響することはない。本研究では、ロジスティック回帰に基づく 2 値分類器をクラスの数だけ用意して、このモデルを構築した。以下では、このモデルを Indep-Logistic と呼ぶ。

Indep-Logistic は単純なモデルであるが、クラスごとに独立に分類器を学習するため、クラス間のある種の相関関係を考慮することができない。ここでいう相関関係とは、例えば、「文学名」に分類されるものの多くは漫画の作品名であり、「番組名」や「映画名」にも分類されやすいといった傾向の事である。本研究では、このような相関関係をラベル付与に取り入れるため、中間層を持つニューラルネットを用いたマルチタスク学習 (Caruana, 1997) を導入する。このモデルでは、図 3b に示すように、クラス数に等しい個数のノードからなる出力層の各ノードで各クラスのラベル付与の確率を出力する。出力層の全てのノードと結合している中間層において、全てのクラスで共有されるパラメタが学習される。これによって、クラス間の何らかの相関関係が学習されることが期待される。以下では、このモデルを Joint-NN と呼ぶ。

Indep-Logistic と Joint-NN の間には、中間層を持つニューラルネットの導入と、マルチタスク学習（中間層の共有）の導入という 2 つの変更がある。これらによるラベル付与の性能の変化を区別するため、実験では、クラスの数だけニューラルネットを構築しそれらを独立に訓練するモデルも構築した（図 3c)。

Indep-Logistic モデルでは、n 次元の素性ベクトル $\boldsymbol{x} \in \mathbb{R}^n$ が与えられた時にラベル c が付与される条件付き確率を以下のようにモデル化した。

$$p_{\text{Indep-Logistic}}(y_c = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}_c \cdot \boldsymbol{x} + b_c) \tag{1}$$

ここに、それぞれのクラス c について、$\boldsymbol{w}_c \in \mathbb{R}^n$ および $b_c \in \mathbb{R}$ は出力層の重みベクトルとバイアス項をそれぞれ示す。

JOINT-NN モデルでは、条件付き確率は以下のようになる。

$$p_{\text{JOINT-NN}}(y_c = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}_c \cdot \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) + b_c) \tag{2}$$

ここに、$\boldsymbol{W} \in \mathbb{R}^{n \times k}$ および $\boldsymbol{b} \in \mathbb{R}^k$ は k 次元の中間層の重み行列とバイアスベクトルをそれぞれ示す。また、それぞれのクラス c について、$\boldsymbol{w}_c \in \mathbb{R}^k$ および $b_c \in \mathbb{R}$ は出力層の重みベクトルとバイアス項をそれぞれ示す。

INDEP-NN モデルでは、条件付き確率は以下のようになる。

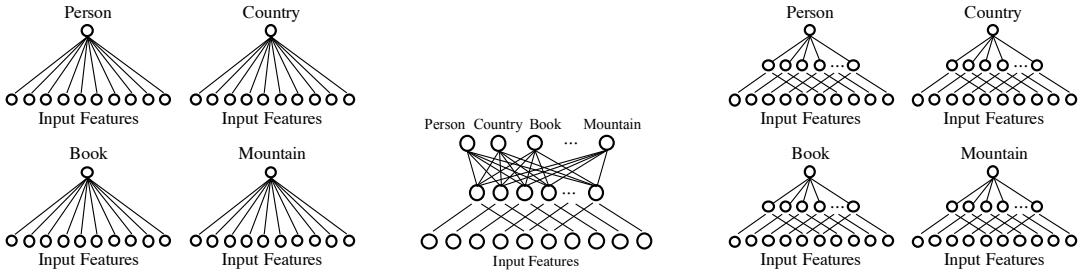
$$p_{\text{INDEP-NN}}(y_c = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}_c \cdot \sigma(\boldsymbol{W}_c\boldsymbol{x} + \boldsymbol{b}_c) + b_c) \tag{3}$$

ここに、それぞれのクラス c について、$\boldsymbol{W}_c \in \mathbb{R}^{n \times k}$ および $\boldsymbol{b}_c \in \mathbb{R}^k$ は k 次元の中間層の重み行列とバイアスベクトルをそれぞれ示す。また、それぞれのクラス c について、$\boldsymbol{w}_c \in \mathbb{R}^k$ および $b_c \in \mathbb{R}$ は出力層の重みベクトルとバイアス項をそれぞれ示す。

それぞれのモデルについて、次式で表される交差エントロピーを損失関数とし、Adam のアルゴリズム (Kingma and Ba, 2014) を用いてそれを最小化した。

$$\mathcal{L} = \sum_{\boldsymbol{x},c} -\{\delta(\boldsymbol{x}, c) \log(p(y_c = 1|\boldsymbol{x})) + (1 - \delta(\boldsymbol{x}, c)) \log(1 - p(y_c = 1|\boldsymbol{x}))\} \tag{4}$$

ここに、$\delta(\boldsymbol{x}, c)$ は、$\boldsymbol{x}$ で表される記事にラベル c が付与されている場合のみ 1 になり、そうでない場合は 0 となる関数である。



(a) INDEP (b) JOINT (c) INDEP-HIDDEN

図 3: ラベル付与の 3 つのモデル

5 素性

ラベル付与のモデルの構築にあたって、2 種類の素性セットを用いた。1 つは既存研究 (Higashinaka et al., 2012) の再現であり、もう 1 つは本研究で提案するものである。

5.1 ベースライン素性

ベースライン素性として、(Higashinaka et al., 2012) で用いられていた素性を可能な限り再現した。表 1 に再現した素性の一覧を示す [1]。

以下では、このベースライン素性を F_b で示す。

5.2 記事ベクトル素性

上に挙げたベースライン素性は、ラベル付与の対象となる記事それ自身の情報をエンコードする上で有効であると考えられる。しかし一方、ラベル付与の対象

[1]元論文 (Higashinaka et al., 2012) で用いられていた素性のうち、T8, T12, T14, M22 で示されていた素性は、内部の資源を用いていたために再現できなかった。また、同様の理由により、形態素解析には JTAG (Fuchi and Takagi, 1998) の代わりに MeCab（http://taku910.github.io/mecab/）を用いた。さらに、Wikipedia から本文を抽出する際には、Wikipedia Extractor（http://medialab.di.unipi.it/wiki/Wikipedia_Extractor）を用いた。

表 1: ベースライン素性の一覧

Features
記事タイトルの単語　unigram
記事タイトルの単語　bigram
記事タイトルの品詞　bigram
記事タイトルの文字　bigram
記事タイトルの最右名詞
記事タイトルの末尾 1 文字
記事タイトルの末尾 3 文字
記事タイトルの末尾 1 文字の文字種
本文 1 文明の最右名詞
記事の見出し名
記事が属する Wikipedia のカテゴリ
記事が属する Wikipedia のカテゴリの上位カテゴリ

の記事が、他の記事からどのような文脈で言及およびリンクされているかといった情報も、記事の分類に重要な情報となりうると考えられる。

例えば、「エベレスト」という記事に固有表現ラベルを付与することを考える。この記事は、他の記事からは次のような文脈でリンクされている。

- … ヒマラヤ山脈の エベレスト の南に連なる …
- … 3 度目の エベレスト 登頂に成功した …
- … 1924 年 エベレスト 遠征隊に参加 …

この例の場合、リンク周辺の「山脈」や「登頂」という語は、リンク先の「エベレスト」という記事に対して固有表現ラベルを付与するにあたり、ラベルを「山地名」とする手がかりになると考えられる。すなわち、記事のリンク元（アンカーテキスト）の文脈を考慮できれば、ラベル付与の性能向上に役立つのではないかと予想される。

リンク元の文脈を表現するには、bag-of-words や 係り受け関係といった、いくつかの方法があるが、本研究では、作られる素性空間のスパースネスの問題に対処するため、Skip-gram (Mikolov et al., 2013a) に基づいて、語の分散表現を学習するという手法をとった。

Skip-gram でリンクの分散表現を学習するにあたって、以下の 3 つの課題が生じた。

- 単純にリンク文字列（アンカーテキスト）を解析の対象としてしまうと、エンティティの曖昧性が生じる場合がある。例えば「ヤマハ」というアンカーテキストからは、「ヤマハ発動機」や（楽器メーカーの）「ヤマハ」といった複数の記事にリンクされているが、アンカーテキストだけではリンク先を一意に定めることはできない。

- Wikipedia の記事名は、「男はつらいよ」のように複数の形態素からなっている場合がある。これらの記事名に対して、単純に形態素解析を適用してしまうと、記事名の途中で区切られてしまい、記事名を 1 語として認識できなくなってしまう。

- 1 つの記事内で、ある他の記事への全ての言及がリンクとしてマークアップされているとは限らない。Wikipedia のガイドラインによれば、同一語に対して全てリンクを貼ることは避けるよう指示されている [2]。

これらの問題に対処するため、以下の工夫を取り入れた。まず、Wikipedia 本文全文に対して、リンクのアンカーテキストをリンク先の記事名に全て置換した。これにより、リンク先の記事の曖昧性が解消される。次に、1 つの記事の中で、少なくとも 1 回はアンカーテキストとして出現した単語は全てリンク先の記事名に置換した。これにより、通常はリンクが貼られない 2 回目以降のエンティティの言及も扱えることになる。これらの処理の過程で、複数の形態素からなる記事名が途中で区切られないように、リンク先の記事名については “<< 男はつらいよ >>” といったようにマークアップすることで、1 語として扱われるようにした。

[2] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

最後に、以上の前処理を施した Wikipedia の全文から、word2vec [3] を用いて単語と Wikipedia 記事名の分散表現（100 次元のベクトル）を獲得した。

以下では、この記事ベクトル素性を F_v で示す.

[3] https://code.google.com/p/word2vec/

表 2: 1 つの記事に付与されるラベル数の分布

付与されたラベルの数	記事数
1	21,624
2	850
3	187
4	14
6	2

6 実験

我々が新たに提案した素性がどの程度有効であるかを評価するために、日本語版 Wikipedia の記事に対して拡張固有表現のラベルを自動的に付与する実験を行った。

6.1 データ

2015 年 11 月 23 日時点の日本語版 Wikipedia より、他の記事からの被リンク数が 100 以上である記事のうちの 22,677 件について、関根の拡張固有表現階層に基づく固有表現分類を人手でアノテートした。Wikipedia には「平和」「睡眠」といった、固有表現ではない事物に関する記事や、「国の一覧」「Wikipedia: 索引」といった、ラベルの付与対象にすべきではない記事がある。それらに対しては、それぞれ「CONCEPT」および「IGNORED」という特別なタグを割り当てることとした。

アノテート済みデータにおける、1 つの記事に付与されるラベル数の分布を表 2 に示す。ほとんどの記事に付与されたラベルは 1 つであったが、4.6% の記事には複数のラベルが付与されていた。

アノテート済みデータにおける、出現頻度が高かった上位 10 ラベルを表 3 に示す。並びに、出現頻度が低かったラベルの例を表 3 に示す。今回は、他記事からの被リンク数が上位の記事をアノテーションの対象としたため、「人名」「番組

表 3: アノテート済みデータにおける出現頻度上位 10 ラベル

ラベル名	記事数	記事の例
人名	4,041	源義経、藤田まこと、ピュートル 1 世
CONCEPT	2,660	国民、ブログ、会社
番組名	2,395	ミュージックフェア、機動新世紀ガンダム X
企業名	1701	日本生命、富士フイルム、会津鉄道
市区町村名	975	東村山市、世田谷区、ロンドン
製品名_その他	964	シンバル、Wii U、Facebook
日付表現	916	5 月 1 日, 2008 年,
文学名	909	フランケンシュタイン、ドラゴンボール、みなみけ
競技会名	625	レスリング世界選手権、札幌オリンピック、菊花賞
IGNORED	621	日本酒の銘柄一覧、2010 年の音楽、2007 年の映画

名」「企業名」といった、日本語版 Wikipedia で参照されやすい記事が多かった一方、「絵画名」「公園名」といった、記事数が少なく、かつ他の記事からの参照も限られるようなラベルの出現は少なくなっていた。

6.2 設定

まず、INDEP-LOGISTIC モデルにおいて、2 種類の素性セット F_b と $F_b + F_v$ を分類器の訓練に用いた場合についてそれぞれ実験を行い、提案手法である記事ベクトル F_v の有効性について検証した。次に、$F_b + F_v$ を分類器の訓練に用いた場合について、2 つのモデル INDEP-NN と JOINT-NN それぞれについて実験を行い、提案手法である、ニューラルネットによるマルチタスク学習の効果を検証した。

INDEP-LOGISTIC モデルについては、L2 正則化ロジスティック回帰分類を行い、実装には scikit-learn(Pedregosa et al., 2011) を用いた。

INDEP-NN モデルと JOINT-NN モデルについては、ニューラルネットの実装に Chainer(Tokui et al., 2015) を用いた。隠れ層の次元は $k = 100$ とし、活性化

表 4: アノテート済みデータにおける出現頻度が低かったラベル

ラベルが付与された記事数	ラベル数	ラベルの例
0	55	URL, 人数, 古墳名, 絵画名
1	8	温泉名, 船名, 恒星名, 両生類名
2-5	16	運河名, 公園名, 橋名, 内閣名
6-10	23	地震名, 条約名, 港名, 惑星名
11-20	23	公共機関名, 昆虫類名, 美術博物館名

関数にはシグモイド関数を用いた。バッチサイズは 10 とした。

それぞれのモデルの訓練時には、データスパースネスや計算時間の問題に対処するため、使用する素性を出現回数が上位の 10,000 種類に限定した。

ラベル付与の性能を評価するために、事例ベースおよびラベルベースの適合率、再現率、F 値を求めた (Godbole and Sarawagi, 2004; Tsoumakas et al., 2009)。

事例ベースの適合率、再現率、F 値は次式で定義される。

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{5}$$

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{6}$$

$$\text{F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{7}$$

ここに、Y_i と Z_i はそれぞれ記事 i の正解ラベルの集合および予測ラベルの集合を表す。N は記事数を表す。

ラベルベースの評価には、通常の適合率、再現率、F 値をラベルごとに求めた。

すべての実験は、10 分割交差検定で行った。

表 5: 事例ベースのラベル付与性能

モデル	Precision	Recall	F1
INDEP-LOGISTIC (F_b)	.8359	.8357	.8334
INDEP-LOGISTIC ($F_b + F_v$)	.8578	.8675	.8583
INDEP-NN	.8707	.8816	.8713
JOINT-NN	**.8853**	**.8862**	**.8831**

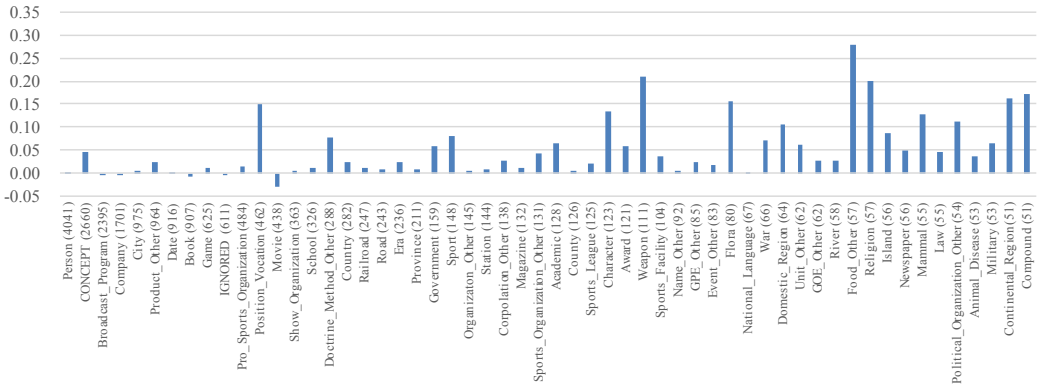


図 4: INDEP-LOGISTIC (F_b) と JOINT-NN を比較した場合の F 値の向上（記事数が 50 以上のクラスのみ, 括弧内は記事数）

6.3 結果

ラベル付与の事例ベースの性能を表 5 に示す。表 5 に示した全ての 2 つの設定の組み合わせについて、ラベル付与性能の向上は二項検定で 1%有意であった。

INDEP-LOGISTIC (F_b) と INDEP-LOGISTIC ($F_b + F_v$) での結果を比較すると、事例ベースの F 値は 2.5 ポイント向上した。

表 6 に、INDEP-LOGISTIC (F_b) と INDEP-LOGISTIC ($F_b + F_v$) を比較してラベル付与の F 値が向上した上位 10 クラスを示す。表 6 に挙げたクラスの多くは、クラス特有の文脈を取りやすいものであったと考えられる。例えば、「植物名」のラベルが付与される記事には、「品種」「花」「咲く」といった語、「宗教名」の記事に対しては「教会」「宗派」「教徒」といった、クラスに特徴的な語がリンク周辺文脈に現れやすいと考えられるが、そのようなクラスのラベル付与性能が、記

表 6: INDEP-LOGISTIC (F_b) と INDEP-LOGISTIC ($F_b + F_v$) とでラベルベースの性能が向上したクラス（記事数が 50 未満のクラスを除く）

ラベル（記事数）	ΔPrecision	ΔRecall	ΔF1
食べ物名_その他 (57)	-0.2229	0.3509	0.1963
宗教名 (57)	-0.1724	0.3158	0.1488
大陸地域名 (51)	-0.0865	0.1961	0.1198
地位職業名 (462)	-0.0553	0.2056	0.1098
武器名 (111)	-0.1419	0.2252	0.1090
哺乳類名 (55)	-0.0231	0.1636	0.0879
植物名 (80)	-0.0398	0.1625	0.0781
単位名_その他 (62)	-0.0186	0.1129	0.0559
主義方式名_その他 (288)	-0.1521	0.1077	0.0553
競技名 (148)	-0.1179	0.1486	0.0461

事ベクトルの導入により向上したと考えられる。

INDEP-LOGISTIC ($F_b + F_v$) と INDEP-NN とでは、1.3 ポイントの F 値の向上が見られた。これは、中間層を持つニューラルネットの導入により、入力素性の組み合わせをラベル付与に用いたことでの性能向上に相当する。

マルチタスク学習の導入によるラベル付与の性能の向上を確認するため、2 つのモデル INDEP-NN と JOINT-NN の間の性能向上を確認した。表 7 に、INDEP-NN と JOINT-NN でラベル付与の F 値が向上した上位 10 クラスを示す。表 7 に挙げたクラスの多くは「＊＊_その他」というクラスであり、また Precision が大きく向上したクラスが多い。関根の拡張固有表現階層において「＊＊_その他」という名前のクラスの多くは、階層におけるその兄弟ノードのクラスに当てはまらないものが分類されるクラスである。例えば「組織名_その他」[4] というクラスには「オックスフォード大学出版局」「NHK 水戸放送局」「新撰組」といった種々雑多なエンティティが分類されるが、これらのクラスの分類性能、特に Precision が

[4]関根の拡張固有表現階層では「組織名の内、その下位のクラスに属さないもの。例えば同好会、クラブなど。また、組織内部につくられた組織（部、課など）」と定義されている。

表 7: INDEP-NN と JOINT-NN とでラベルベースの性能が向上したクラス（記事数が 50 未満のクラスを除く）

ラベル（記事数）	ΔPrecision	ΔRecall	ΔF1
化合物名 (51)	0.1058	0.0784	0.0909
組織名_その他 (145)	0.1296	0.0483	0.0782
政治的組織名_その他 (54)	0.2158	0.0000	0.0771
競技組織名_その他 (131)	0.0394	0.0763	0.0604
キャラクター名 (123)	0.0981	0.0326	0.0564
文学名 (907)	0.0572	0.0484	0.0526
GPE_その他 (85)	0.0706	0.0353	0.0498
法人名_その他 (138)	0.1368	-0.0072	0.0489
島名 (56)	0.1012	0,0000	0.0486
武器名 (111)	0.1249	-0.0181	0.0471

向上したということは、「＊＊_その他」の兄弟のクラスとの相関が学習され、余計な記事が「＊＊_その他」に分類されなくなったためではないかと考えられる。

提案手法による最終的なラベル付与性能の向上を調べるため、INDEP-LOGISTIC (F_b) と JOINT-NN の間のラベルベースの F 値の変化クラスごとに求めた。図 4 は、それらを記事数の多いラベルから順に並べたものである。図 4 より、提案手法によって、特に記事数の少ないクラスについてラベル付与の性能が大きく向上したことがわかる。

個別の事例を確認すると、ラベル付与の閾値を変化させることで改善が可能とみられる事例が幾つか見つかった。実際に JOINT-NN での誤りの個数を数えてみると、予測ラベルが正解ラベルと完全一致しなかった 3195 件のうち 1328 件 (41.6%) は、今回用いた閾値の設定では予測ラベルとして 1 つもラベルを付与できていなかった。これは、ラベル付与の閾値を変更することで解消される問題であると考えられるが、今後の調査が必要である。それに加えて、必ずしも予測ラベルが誤りとは言えないような例や、人手でアノテーションしたラベルが誤っていた例もいくつか存在した。今回は人手でのラベル付与を 1 人によって行っていた

表 8: Joint-NN での誤り

記事名	予測ラベル	アノテートされた正解ラベル
Twitter	CONCEPT; 製品名_その他	製品名_その他
米	植物名; 食べ物名_その他	植物名
マーシャル諸島	国名; 島名	国名
K-1	競技会名	競技リーグ名
酵素	CONCEPT	自然物名_その他
二条城	遺跡名_その他	神社寺名
ちはやふる	番組名; 文学名; 映画名	番組名

が、今後は複数人でラベル付与を行い、作業者間での一致率をみてアノテーションの信頼性や妥当性を保証することが必要になると考えられる。

7　おわりに

本稿では、Wikipedia の記事に対して、細かい粒度の固有表現のラベルを付与するタスクに取り組んだ。分類の粒度を細かくすることによって生じる、項目数が少ないクラスに対するデータスパースネスの問題に対処するため、すべてのクラスの分類を同時に学習するマルチタスク学習を導入し、これを中間層を持つニューラルネットワークによって実現した。これにより、特に項目数の少ないクラスにおいて分類の性能が向上した。また、分類器の構築に用いる素性として、従来よく用いられていた、記事の内容語の bag-of-words のような離散的な情報のみでは素性空間が疎になるという問題に対して、我々は Skip-gram モデルに基づく手法によって、記事のリンク元の文脈を反映した連続的な分散表現を獲得し、分類器の構築の素性の一部として用いたことで、離散的な素性のみを用いた場合と比較して、分類の性能が全体的に向上することを示した。

本稿で提案した手法は、言語にもオントロジーにも依らず適用可能なものである。今後の課題として、異なる言語やオントロジーでの本手法の適用についても取り組みたい。

謝辞

本研究を進めるにあたり、ご指導をいただいた乾健太郎教授、岡崎直観教授に感謝いたします。そして、データの提供ならびに実験や論文執筆にあたっての直接の指導をくださった関根聡氏と研究員の松田耕史氏に感謝いたします。最後に、日常の議論を通じて多くの知識や指摘をくださった乾・岡崎研究室の皆様に感謝いたします。

参考文献

Aprosio, A. P., Giuliano, C., and Lavelli, A. (2013). Extending the coverage of DBpedia properties using distant supervision over Wikipedia. In *Proceedings of ICON 2013*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *Proceedings of ISWC'07/ASWC'07*.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Chang, J., Tzong-Han Tsai, R., and S. Chang, J. (2009). Wikisense: Supersense tagging of wikipedia named entities based wordnet. In *Proceedings of PACLIC 23*.

Dakka, W. and Cucerzan, S. (2008). Augmenting wikipedia with named entity tags. In *Proceedings of 3rd IJCNLP*.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ace) program tasks, data, and evaluation. In *Proceedings of LREC 2004*.

Fuchi, T. and Takagi, S. (1998). Japanese morphological analyzer using word co-occurrence - jtag. In *Proceedings of ACL '98 and Proceedings of COLING '98*.

Godbole, S. and Sarawagi, S. (2004). *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings*, chapter Discriminative Methods for Multi-labeled Classification, pages 22–30. Springer Berlin Heidelberg.

Higashinaka, R., Sadamitsu, K., Saito, K., Makino, T., and Matsuo, Y. (2012). Creating an extended named entity dictionary from wikipedia. In *Proceedings of COLING 2012*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *ICLR 2015*.

Ling, X., Singh, S., and Weld, D. S. (2015). Design challenges for entity linking. *TACL 2015*, pages 315–328.

Mann, G. S. (2002). Fine-grained proper noun ontologies for question answering. In *Proceedings of SEMANET '02*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sekine, S., Sudo, K., and Nobata, C. (2002). Extended named entity hierarchy. In *Proceedings of LREC 2002*.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the WWW 2007*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Tardif, S., Curran, R. J., and Murphy, T. (2009). Improved text categorisation

for wikipedia named entities. In *Proceedings of ALTA Workshop 2009*, pages 104–108.

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Toral, A. and Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of Workshop on New Text, EACL 2006*.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Watanabe, Y., Asahara, M., and Matsumoto, Y. (2007). A graph-based approach to named entity categorization in wikipedia using conditional random fields. In *Proceedings of EMNLP-CoNLL 2007*.