Wikipedia 記事に対する拡張固有表現ラベルの多重付与

鈴木 正敏 † 松田 耕史 ‡ 関根 聡 § 岡崎 直観 ‡ 乾 健太郎 ‡
† 東北大学工学部 ‡ 東北大学大学院情報科学研究科 § ランゲージ・クラフト
{m.suzuki,matsuda,okazaki,inui}@ecei.tohoku.ac.jp sekine@languagecarft.com

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．

Wikipedia は，最も大規模なオンライン百科事典である．誰でも閲覧や編集に参加することができ，扱われる内容が広範で，更新も活発であることから，言語資源として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，個々のエンティティに対して「人名」「地名」などといった固有表現クラスのラベルに関する知識を構築することが重要になる．固有表現クラスは，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エンティティ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．

また，エンティティの固有表現クラスについての大規模で詳細な知識は，それ単体でも自然言語処理の様々なタスクにおいて有用であることが知られている．Ling ら [7] は 固有表現抽出やエンティティリンキングの性能向上に役立つことを指摘している．

Wikipedia の情報を構造化して知識ベースを構築する試みの先駆的なものとして DBpedia [3] と YAGO [10] がある．しかしながら，これらはいずれもオントロジと記事の対応付けに人手の努力やメタデータに基づいた単純なルールを用いており，メタデータが乏しい記事や，新しいクラスへの対応，カバレッジの点で課題が残る．また，それら知識ベースがクラス体系として用いているオントロジも，コミュニティベースでボトムアップに設計・維持されていたり，WordNet などの異なるドメインを対象とするオントロジを下敷きとしたものが多く，粒度や均一さの面で課題がある [13].

そこで，我々は，集中的にコントロールされた固有表現のクラス階層として**関根の拡張固有表現階層** [1] に着目し，Wikipedia の記事に対して拡張固有表現のラベルを機械学習により自動で付与することに取り組む．本研究では，機械学習によるラベル付与において，どのような情報を素性に用いることが有効であるか，および，適切なラベルを記事に付与する上でどのようなタスク設計が望ましいかを検討し，実験を行った．

[1] https://sites.google.com/site/extendednamedentityhierarchy/

本研究の貢献は，以下の 2 つである．

- 機械学習によって Wikipedia の記事に拡張固有表現のラベルを付与する上で有効な素性をあらためて検討，設計した．特に，記事のリンク元文脈から学習した記事の分散表現を学習の素性として用いることで，ラベル付与の適合率と再現率がともに向上することを示した．
- それぞれのクラスのラベル付与を独立に学習するのではなく，すべてのクラスのラベル付与を同時に行う系をニューラルネットワークによって構築することで，ラベル付与の性能がさらに向上することを示した．特に，訓練データに含まれる文書数が比較的少ないクラスでは分類性能の大きな改善が見られた．

本研究のタスクの概観図を図 1 に示す．

2 関連研究

Aprosio ら [2] は，言語間リンクを用いて自動的に生成した教師データを利用して分類器を訓練し，DBpedia の知識を自動的に拡張する試みについて述べているが，他の言語の記事から分類に必要な情報が得られる状況は限られている．また，Alotaibi ら [1] や Nothman ら [9] は Wikipedia の記事を固定された固有名詞クラスに分類することを試みているが，彼らが対象としているクラス体系は 10 クラス程度の荒い粒度のものである．

Wikipedia の記事に対して，関根の拡張固有表現のクラスを割り当てる試みもいくつか存在する．たとえば，Higashinaka ら [6] は，記事のタイトル，本文の 1 文目，記事が属するカテゴリといった情報から，ラベル付与に有効な素性を抽出し，教師あり学習による分類問題として解いた．また，杉原ら [12] は，Wikipedia におけるカテゴリ間の関係に着目した教師あり学習に基づく分類手法を提案している．これらの研究は共通して，それぞれの項目に拡張固有表現のラベルを 1 つだけ付与するシングルラベル分類としてタスクを設計しているため，一部の記事が持つ多義性に対処できないという課題がある．また，学習で用いる素性として，記事のリンク元の文脈や，「一覧記事」の内容といった，Wikipedia が特徴的に備えている情報を利用することも可能であるはずだが，そのような情報はあまり活用されていない．

3 タスクの設計

3.1 マルチラベル分類

文書や記事に対してラベルを付与するタスクは，1 つの文書につきラベルを 1 つだけ付与するシングルラベル分類と，1 つの文書に対して複数のラベルの

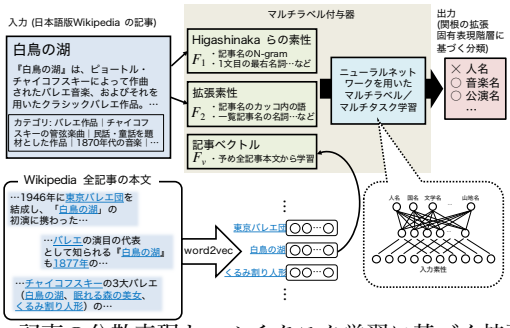


図 1: 記事の分散表現とマルチタスク学習に基づく拡張固有表現ラベルの自動付与

付与を認めるマルチラベル分類の 2 つに大別されるが，本研究が対象とする問題は，次の例からも分かるようにマルチラベル分類と捉えるべきである．

> **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，記事が複数のカテゴリに属す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る．

3.2 マルチタスク学習

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（**シングルタスク学習**）．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．

しかし実際には，クラスとクラスの間には，何らかの相関関係があると考えられる．例えば,「文学名」のラベルが付与される記事は，同時に「映画名」のラベルも付与される場合が多い一方,「道路名」のラベルは付与されることは少ない．このようなクラス間の相関関係を，学習の段階でクラス間で共有することができれば，ラベル付与の性能向上につながる可能性がある（**マルチタスク学習**）[4]．また，シングルタスク学習では学習が難しい，出現頻度の低いクラスにおいても，マルチタスク学習を導入することで，事例数が十分な他のクラスから学習されたパラメータを利用できるようになり，ラベル付与の性能が向上する可能性がある．

そこで本研究では，ニューラルネットワークを用いたマルチタスク学習をラベル付与器の学習において導入し，ネットワークの隠れ層においてクラス間で共有される表現を獲得することを試みた．マルチタスク学習によって，シングルタスク学習と比較して，ラベル付与の性能がどう変化するかを評価した．

4 ラベル付与器の構成

4.1 学習で用いる素性の検討

素性設計のベースラインとして，記事名の単語 unigram / bigram や本文 1 文目の最右名詞，および記事が属するカテゴリ名の最右名詞といった，Higashinaka らの手法 [6] を部分的に再現したもの[2]（以降，F_1 で表す）を用いた．

さらに，Wikipedia が特徴的に備える情報を積極的に利用するため，ならびに固有表現と一般表現の識別を行うために，以下の素性を新たに加えた（以降，F_2 で表す）．

- 記事タイトルのカッコ内の語 1-gram（Bag-of-Words）[3]
- 本文 1 文目のすべての名詞（Bag-of-Words）
- 記載がある一覧記事のタイトルに出現する名詞（Bag-of-Words）
- Wiktionary へのリンクが存在するか（真偽値）
- 岩波国語辞典の見出し語との完全一致があるか（真偽値）[4]

4.2 記事の分散表現の学習

Wikipedia の記事は，一般的な文書と異なり，記事同士が本文中で相互にリンクされている．例えば,「エベレスト」という記事は，他の記事から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，リンク先の「エベレスト」という記事に対してラベルを付与するにあたり，周辺の「山脈」や「登頂」という語は，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．

文脈の表現方法としては，リンクの周辺に出現する単語の Bag-of-Words や，アンカーテキスト周辺の係り受け情報などが考えられるが，本研究では，データスパースネスを考慮し，アンカーテキスト周辺の単語から word2vec [8] を用いて学習した，リンク先の記事の分散表現（200 次元のベクトル）を機械学習の素性（以降，F_v で表す）として用いた．具体的には，Wikipedia 全記事の本文からアンカーテキストをリンク先記事名に置換し，それ以外の部分は形態素ごとに区切って word2vec の入力とした．

4.3 ラベル付与器の構成

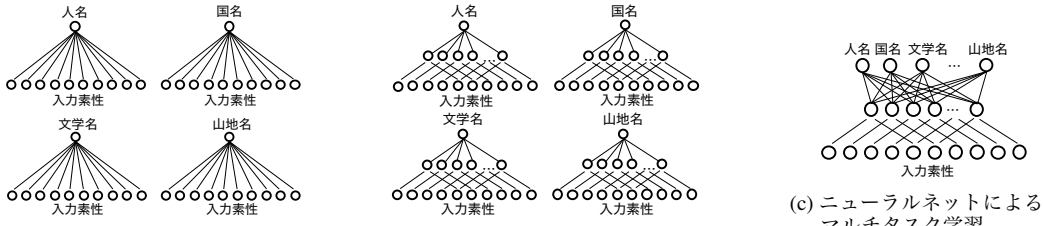
シングルタスク学習およびマルチタスク学習によるラベル付与の実験をする上で，図 2 のような 3 種類のラベル付与器を構成した．

図 2a の構成 SINGLE-LOGISTIC は，既存研究 [6][12] と同様に，クラスの数だけ 2 値分類器を作り，それ

[2]元論文 [6] で評価実験に用いられていた 15 の素性のうち，T8, T12, T14, M22 は，今回の実験では省略した．また，形態素解析器には `MeCab` を用い，記事本文の抽出には `WikiExtractor` を用いている．

[3]地名の「ダウンタウン（ロサンゼルス）」，タレントの「ダウンタウン（お笑いコンビ）」のように，同名の記事の曖昧性を解消するためのカッコの内部の語である．

[4]『岩波国語辞典第五版タグ付きコーパス 2004』を使用した．



(a) ロジスティック回帰による
シングルタスク学習
SINGLE-LOGISTIC

(b) ニューラルネットによる
シングルタスク学習
SINGLE-NN

(c) ニューラルネットによる
マルチタスク学習
MULTI-NN

図 2: 構成した 3 種類のラベル付与器

ぞれのクラスのラベルを付与すべきかどうかをクラスごとに独立に学習する，シングルタスク学習の構成である．

図 2b の構成 SINGLE-NN は，SINGLE-LOGISTIC と同様にシングルタスク学習であるが，それぞれの分類器を隠れ層 1 層のニューラルネットワークとした．隠れ層を導入することによって，シングルタスク学習のままでもラベル付与の性能が向上するかどうかを調査するためにこの条件を設定した．

図 2c の構成 MULTI-NN は，ニューラルネットワークを用いたマルチタスク学習である．SINGLE-LOGISTIC および SINGLE-NN のシングルタスク学習と異なり，各クラスが共通の隠れ層を持っている．この共通の隠れ層を導入し，すべてのクラスを同時に学習することで，各クラスのラベル付与は独立ではなくなり，必要な情報がクラス間で共有されることが期待される．

5 実験

5.1 データ

実験に用いるために，日本語版 Wikipedia の記事（2015 年 11 月 23 日版）で，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 22677 件について，関根の拡張固有表現階層（ver. 7.1.0）に基づく固有表現分類ラベルを人手で複数付与した．

なお，Wikipedia には「平和」「睡眠」といった固有表現ではない一般的な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．

このデータの構築方法やラベルごとの事例数などの詳細は，関根らによる報告 [13] を参照されたい．

5.2 設定

まず，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，学習で用いる素性セットが F_1，$F_1 + F_2$，$F_1 + F_2 + F_v$ の 3 通りの場合について，SINGLE-LOGISTIC の構成で実験を行った．それぞれの分類器の出力確率の閾値は 0.5 とした．また，それぞれの素性セットについて，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 10000 種類に限定した．

次に，構成をニューラルネットを利用した SINGLE-NN および MULTI-NN とすることでラベル付与の性能がどのように変化するかを評価する実験を行った．学習に使用する素性セットとして $F_1 + F_2 + F_v$ を用い，前の実験と同様に出現回数が上位 10000 種類の素性を用いた．

ラベル付与の性能を評価する指標として，事例ベースの適合率，再現率，F 値，およびクラスベースの適合率，再現率，F 値を用いた [5][11].

事例ベースの適合率，再現率，F 値は以下の式で表される．

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

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

$$F_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{3}$$

ここで，Y_i，Z_i はそれぞれ事例 i の正解ラベル集合，予測ラベル集合である．

クラスベースの評価では，各クラスのラベルについて通常の適合率，再現率，F 値を求めた．

すべての実験は，10 分割交差検定で行った．

5.3 結果と考察

実験結果を表 1 に示す．SINGLE-LOGISTIC では，追加した素性によって，適合率と再現率の両方で改善が見られた．

MULTI-NN では，さらにラベル付与の性能が向上し，事例ベースの F 値がベースラインの SINGLE-LOGISTIC と比較して最大でおよそ 2 ポイント向上した．

SINGLE-NN の分類性能が SINGLE-LOGISTIC と MULTI-NN のおよそ中間となったことから，隠れ層の導入とマルチタスク学習のどちらも，ラベル付与の精度向上に貢献したと考えられる．MULTI-NN では，隠れ層の次元数や層数を増やすことで性能の変化があるかどうかも実験したところ，わずかながら性能の向上が見られた．

記事数が少ないクラスについてラベル付与の性能がマルチタスク学習によってどう変化したかを調べるために，ラベル付与器の学習方法を SINGLE-LOGISTIC（素性: $F_1 + F_2 + F_v$）から MULTI-NN（隠れ層 1 層（200 次元））に変更したことによるクラスごとの F 値の変化を求めた．図 3 は，それらを記事数の多いクラスから順に並べたものである．この図から，記事数の少ないクラスほどラベル付けの性能がより向上する傾向が見られ，マルチタスク学習がラベル付与の頑健性に大きく貢献していることが分かった．

一方，複数のラベルを付与すべき項目についてラベル付与の性能が向上したかどうかを同様に調べる

表 1: シングルタスク学習およびマルチタスク学習によるラベル付与の結果（事例ベース）

ラベル付与器の構成	条件	適合率	再現率	F 値
SINGLE-LOGISTIC	素性: F_1	.8337	.8337	.8314
	素性: $F_1 + F_2$	.8501	.8524	.8486
	素性: $F_1 + F_2 + F_v$	.8647	.8720	.8647
SINGLE-NN	隠れ層 1 層（100 次元）	.8701	.8783	.8701
MULTI-NN	隠れ層 1 層（100 次元）	.8820	.8810	.8792
	隠れ層 1 層（200 次元）	.8860	.8868	.8839
	隠れ層 2 層（100 次元 2 つ）	**.8863**	**.8871**	**.8842**

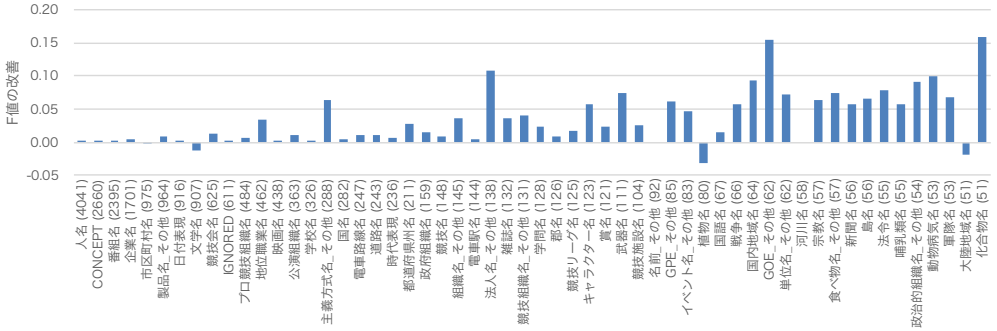


図 3: マルチタスク学習の導入によるクラスベースの F 値の改善（記事数が 50 以上のクラスのみ，括弧内は記事数）

表 2: マルチタスク学習の導入による複数ラベルの記事の F 値の変化（F_S と F_M はそれぞれ SINGLE-LOGISTIC と MULTI-NN での F 値を表す）

ラベル (記事数)	F_S	F_M	$F_M - F_S$
番組名; 文学名 (496)	.7237	.6915	-0.0363
番組名; 映画名; 文学名 (80)	.4872	.5309	0.0437
番組名; キャラクター名 (62)	.7568	.7797	0.0229
番組名; 製品名_その他; 文学名 (60)	.2609	.2314	-0.0295
製品名_その他; 文学名 (41)	.2078	.1905	-0.0173
食べ物名_その他; 植物名 (36)	.7826	.7536	-0.0290
番組名; 組織名_その他 (34)	.8060	.8286	0.0226
映画名; 文学名 (33)	.3333	.2727	-0.0606

表 3: MULTI-NN でのラベル付与の誤りの例

記事名	予測ラベル	正解ラベル
東京スポーツ	新聞名	企業名
日本郵便	企業名	法人名_その他
K-1	競技会名	競技リーグ名
酵素	CONCEPT	自然物名_その他
ちはやふる	番組名; 文学名; 映画名	番組名

と，表 2 に示すように，記事数の多少にかかわらず性能の向上は見られず，マルチタスク学習によってラベル間の相関を捉えることによる効果は今回の実験からは確認できなかった．

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 MULTI-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 3154 件のうち 1394 件（44.2%）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，クラスごとに異なる値を持つであろうバイアス項の重みが，シングルタスク学習の場合は各クラスに対して独立に学習されるのに対し，マルチタスク学習においては全クラスから学習されることに起因する問題であると推測されるが，今後も調査が必要である．その一方で，表 3 に示すように，誤り事例の中には，必ずしも誤りと言えないものも多く，人手でのラベル付与にまだ改善の余地が残されていることを示唆している．

6 おわりに

本研究では，Wikipedia の記事に拡張固有表現ラベルを付与するにあたり，記事のリンク元文脈を学習で用いる素性として取り入れること，ならびにマルチタスク学習の導入がラベル付与の性能の向上につながることを示した．

Wikipedia の個々の記事に対して拡張固有表現のラベルを適切に付与できれば，記事から抽出すべきエンティティの属性やエンティティ同士の関係が明確になり，Wikipedia 上の知識の構造化を次のステップへ進めることが出来る．今後は，固有表現ラベルが付与された記事からの情報抽出や記事同士の関係抽出などのタスクに取り組みたい．

謝辞

この研究は，文部科学省受託研究「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」および，文部科学省科研費（15H01702, 15H05318）の一環として行われた．ならびに，データを作成していただいたランゲージ・クラフトの安藤様に感謝いたします．

参考文献

[1] Fahd Alotaibi and Mark Lee. Mapping Arabic Wikipedia into the Named Entities Taxonomy. In COLING, 2012.
[2] Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information. In 10th Extended Semantic Web Conference, 2013.
[3] Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In 6th Int'l Semantic Web Conference, 2007.
[4] Rich Caruana. Multitask learning. In Machine learning, pp. 41-75, 1997.
[5] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In PAKDD, 2004.
[6] Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, Yoshihiro Matsuhiro. Creating an Extended Named Entity Dictionary from Wikipedia. In COLING, 2012.
[7] Xiao Ling, Sameer Singh, and Daniel S. Weld. Design Challenges for Entity Linking. In TACL, 2015.
[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In NIPS 2013.
[9] Joel Nothman, James R. Curran and Tara Murphy. Transforming Wikipedia into named entity training data. In Australian Language Technology Workshop. 2008.
[10] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago - A Core of Semantic Knowledge. In WWW, 2007.
[11] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining Multi-label Data. In Data Mining and Knowledge Discovery Handbook, Springer, 2010.
[12] 杉原大悟, 増市博, 梅本宏. Wikipedia カテゴリ階層構造の固有名詞分類実験における効果. 情報処理学会研究報告. 情報学基礎研究会報告（NL-189-9）, pp. 57-64, 2009.
[13] 関根聡, 安藤まや, 松田耕史, 鈴木正敏, 乾健太郎. 「拡張固有表現＋Wikipedia」データ. 言語処理学会第 22 回年次大会, 2016.