Vol.2018-NL-238 No.14
2018/12/12

文情報の活用と階層構造に着目した固有表現ラベル推定手法

白水 優太朗[1,a)]　藤本 拓[1]　吉村 健[1]　礒田 佳徳[1]

概要：語の固有表現ラベルを推定することは，スロット・フィリングや上位概念語辞書作成などにおいて有用である．しかし，従来の手法では，推定対象の語が限定されていたり，十分な量の特徴量抽出が難しいという問題点がある．本稿では，容易かつ大量に獲得可能な入力素性として，語が含まれる文情報及び語の文字情報を利用し，さらに，固有表現ラベルの各階層を同時に推定する手法を提案する．実験により，従来手法よりもラベルの推定精度が向上することを示す．

Shiramizu Yutaro[1,a)]　Fujimoto Hiroshi[1]　Yoshimura Takeshi[1]　Isoda Yoshinori[1]

1. はじめに

語の固有表現ラベルを推定することは，情報抽出や質問応答など，自然言語処理の幅広い分野の要素技術として重要である．例えば対話システムにおいて，「電車駅名から電車駅名まで行きたい」という発話パターンを乗換案内タスク[*1]に割り当てたとき，「渋谷駅」「品川駅」といった語の固有表現ラベルが「電車駅名」であることが分かれば．「渋谷駅から品川駅まで」というユーザ発話の意図を乗換案内タスクだと推定できる．このようなケースでは，語と固有表現ラベルのペアを収録した上位概念語辞書（固有表現ラベル辞書）を事前に作成しておくことが効果的である．語に対して推定・付与される固有表現ラベル体系についても，PERSON や LOCATION といったレベルの粗粒度のもの [4][12] から，100 種類を超える細粒度のもの [6][8][13] まで多く提案されている．

以上の背景を踏まえて，語に対し適切な固有表現ラベルを推定する手法が盛んに研究されている [2][5][10][14][15][16][18]．しかし，多くの研究は Wikipedia の記事名（Wikipedia エンティティ）を対象としたラベル推定であり，Wikipedia に由来する素性を特徴量として利用している．このため，Wikipedia に収録されていない新語・未知語には適用できないという問題がある．また，階層的な固有表現体系を利用しているにもかかわらず，最下層の固有表現ラベルのみを語に対して推定・付与しており，ラベルの階層構造を十分に活用できているとはいえない．

本研究は，容易に大量獲得が可能な Wikipedia に拠らない入力特徴量として，語が含まれる文情報及び語の文字情報を利用することで，Wikipedia に未収録の語に対するラベル推定を実現する．文情報は文中の語のユニグラムを利用し，語の文字情報は語の文字ユニグラムを利用する．また，固有表現ラベルの階層構造に着目し，階層内・階層間の情報を共有したモデルを提案する．さらに，提案する特徴量および提案モデルが実際の固有表現ラベル推定において有効であることを，「拡張固有表現＋ Wikipedia」データ [19] に対して確認する．

本研究の貢献は下記 2 点である．

- 文情報及び文字情報を特徴量として利用することで，分類性能が向上することを示した．文は大規模獲得が比較的容易であり，先行研究と比べて特徴量作成コストを削減できた．
- 各階層の固有表現ラベルを同時に推定することで分類性能が向上することを示した．

2. 関連研究

Wikipedia 記事名（Wikipedia エンティティ）に対し固有表現ラベルを推定する研究は多く挙げられる [2][10][14]．これらの研究は，Wikipedia 記事名に対し，3–15 程度の粗粒度の固有表現ラベルを付与する問題に取り組んでいる．

[1] 株式会社 NTT ドコモ
NTT DOCOMO, INC., 4-5, Akasaka 2-chome, Minato-ku, Tokyo 107–0052, Japan
[a)] yuutarou.shiramizu.pf@nttdocomo.com

[*1] ユーザの発話を解析し，ユーザの求めているタスクが「乗換案内」であると判定すること

しかし，粗粒度の固有表現ラベルでは，対話システムなど他領域への活用を検討したときに問題が生じる．例えば，抽出タスクにおいてよく使われる分類体系のひとつであるIREX[19] において，「アメリカ合衆国」「赤坂見附」「太平洋」という語にはすべて LOC（「地名」）が付与されるが，それぞれ「国名」「国内地域名」「海洋名」といった，細かい粒度の固有表現ラベルが付与された方がより有用であると考えられる．

細粒度の固有表現ラベルを推定する試みとして，Higashinaka ら [5] は，ラベル推定対象の Wikipedia 記事の「記事本文 1 文目に含まれる名詞」や，「記事に付与されている Wikipedia カテゴリ」といった Wikipedia 固有の情報から，ラベル推定に有効な特徴量を抽出し，Wikipedia 記事名に対して拡張固有表現階層 [13] を推定する分類器を構築している．さらに，各特徴量の効果を比較検討し，ラベル推定に有用な特徴量について報告している．また，杉原ら [15] は，Wikipedia から Wikipedia カテゴリ間のリンク構造を抽出し，主要カテゴリからの最短経路が固有表現ラベル推定（特に Recall の向上）に効果的であることを示している．

上記の研究はいずれもシングルラベル分類問題としてのアプローチであり，1 つの Wikipedia 記事名に対し 1 つの固有表現ラベルを推定している．しかし，例えば「バナナ」という記事名には「植物名」と「食べ物名_その他」といった複数の固有表現ラベルを付与できるように，語が持つ多義性を考慮すると，固有表現ラベル推定タスクをシングルラベル分類問題と捉えるのはふさわしくない．

鈴木ら [18] は，1 つの記事名に対し複数のラベルの推定を認めるマルチラベル分類問題として，Wikipedia 記事名に対する固有表現ラベル推定を行っている．さらに，固有表現ラベル間には相関があると仮定し，ラベル推定にマルチタスク学習 [1] を適用している．すなわち，ある語（例えば「吾輩は猫である」）に対して固有表現ラベル「文学名」が推定されるとき，同時に「映画名」も推定されやすく，一方で「道路名」は推定されにくいといったラベル間の相関関係を踏まえ，隠れ層を共有したニューラルネットワークによって全ての固有表現ラベルを同時に学習・推定する手法を提案している．固有表現ラベルは，Higashinaka ら [5] や杉原ら [15] と同様に，拡張固有表現階層 [13] を利用している．特徴量は，Higashinaka ら [5] によって用いられたものを部分的に再現しただけでなく，「Wikipedia エンティティの分散表現」や，「一覧記事[*2]の内容」など，Wikipedia が備えている特徴を積極的に活用している．

これらの研究は，Wikipedia 記事名に対する固有表現ラベルの推定を対象としており，特徴量も，「記事本文 1 文目で最後に出現する名詞」や「記事が属するカテゴリ名」など，Wikipedia から抽出できるものを利用している．このため，Wikipedia からの特徴量抽出ができない Wikipedia に未収録の新語や商標に対して固有表現ラベルを推定できないという共通の問題点がある．スロット・フィリングや上位概念語辞書作成などへの応用を考えると，Wikipedia に収録されている語彙だけでは不十分であるため，Wikipedia に由来しない特徴量を用いることで任意の語に固有表現ラベルを推定できる手法の方が望ましい．

水木・榊 [16] はこの問題に対処するため，ラベル推定対象の語の類語の上位語の集合である疑似上位語セットを利用して，Wikipedia の見出し語に対する固有表現ラベルの推定手法を提案している．例えば，「ソフィア」という語の固有表現ラベルを推定する場合，「ソフィア」の類語である「ロンドン」「ブカレスト」「ザグレブ」といった語の上位語（それぞれ「首都」「都市」「都市」）を大規模テキストから収集し，「ソフィア」の疑似上位語セットとして利用する．この疑似上位語セットは自動獲得のため，必ずしも正しいとはいえない上位語がノイズとして入ってくる可能性があるが，Self-Attention Mechanism[7] を用いて，この影響を低減している．また，鈴木ら [18] と異なり，固有表現ラベルと付与ラベル数の推定を同時に行うアーキテクチャを提案している．

水木・榊の提案手法は，Wikipedia に由来しない特徴量を用いており，Wikipedia に未収録の語にも適応できる点で，従来手法よりも汎用性が高い．しかし，上位概念語の獲得に工夫が必要なため，特徴量の選定には未だ検討の余地があるといえる．そこで本研究では，語が含まれる文（テキスト）情報及び語の文字情報を利用して固有表現ラベルを推定する手法を提案する．文は，Web クローリングなどを活用することで大規模に収集でき，疑似上位語よりも容易に獲得できる特徴量だと考えられる．さらに，鈴木ら [18] の研究を踏まえ，同一階層内のラベル間だけでなく階層間にも相関があるという仮定の下，各階層を同時に推定するネットワークアーキテクチャを提案し，その有効性を検証する．

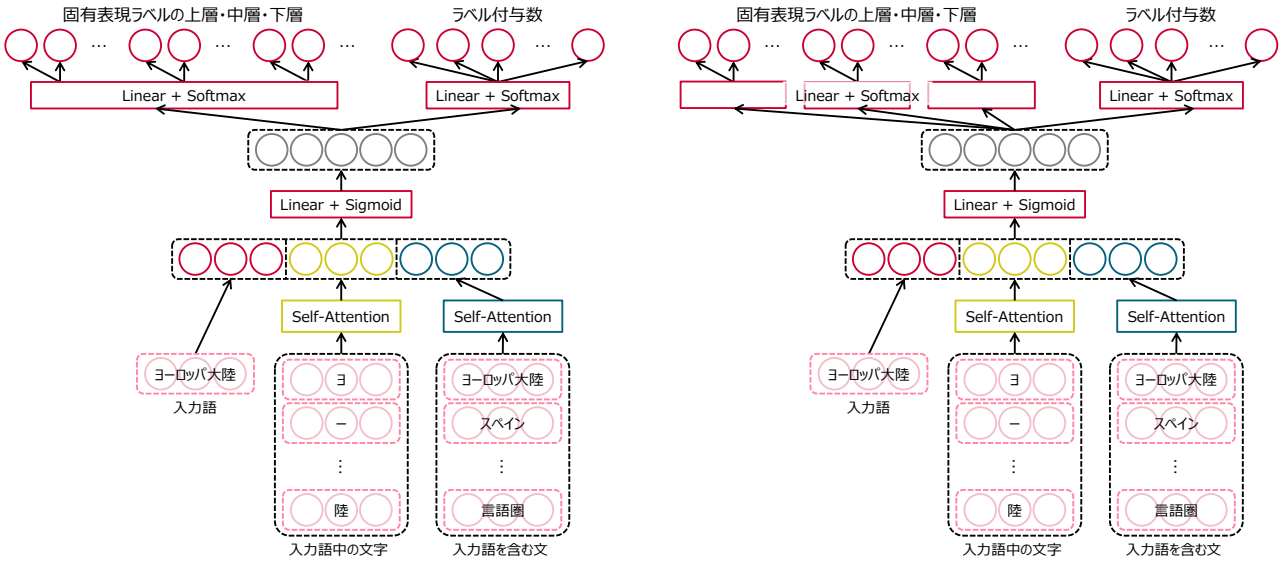
3. 提案手法

3.1 特徴量設計

本研究では，モデルの入力特徴量として，ラベル推定対象の語が含まれる文，ラベル推定対象の語の文字ユニグラム，ラベル推定対象の語を，それぞれ分散表現に変換して用いた．

文は，Wikipedia 本文を，形態素とアンカーテキスト（リンク元の語）の両方で分割したものを利用した．「…ソシュール以降の言語学は言語を…」という文（下線部は「フェルディナン・ド・ソシュール」へのアンカーテキスト）を例とした，具体的な分割手順を図 2 に示す．一般的な形態素解析器による分かち書き（①）に加え，2 個から

[*2] 「国の一覧」や「ノーベル賞受賞者の一覧」など，何かしらの基準に従って物事を列挙した記事．



(a) 各階層の固有表現ラベルを同時に推定　　(b) 各階層ごとに固有表現ラベルを推定

図 1: 提案モデル

…ソシュール以降の言語学は言語を…

① 一般的な形態素解析器による分かち書き

…ソシュール/以降/の/言語/学/は/言語/を…

② 最長一致での検索

…ソシュール/以降/の/言語学/は/言語/を…

③ アンカーテキストの置換

…フェルディナン・ド・ソシュール/以降/の/言語学/は/言語/を…

④ 不要な語の削除

…フェルディナン・ド・ソシュール/言語学/言語…

図 2: 文の分割の例（青文字はアンカーテキスト）

N 個までの形態素を連結した語を Wikipedia 本文中に出現するアンカーテキストで検索し，最長一致でマッチしたアンカーテキストに置換した（②「言語」「学」を「言語学」へ置換）．さらに，本文に元々あったアンカーテキストは，リンク先の記事名へ置換した（③「ソシュール」を「フェルディナン・ド・ソシュール」に置換）．それ以外の部分は，形態素ごとに区切り，動詞や形容詞，記号，機能語などは削除した（④）．通常の形態素解析ではなく上記の分割手順を踏むことのメリットとして，形態素の情報だけではなく，Wikipedia エンティティや Wikipedia エンティティに関する語（アンカーテキスト）の情報を十分に捉えられる点が挙げられる．

本研究では $N = 10$ とし，形態素解析には JTAG[3] を用いた．また，上記プロセスで分割した文を word2vec[9] に入力し，分散表現（300 次元）を学習した．

文及び文字ユニグラムは，先行研究 [16] と同様に Self-Attention Mechanism[7] を利用し，文中・語中に含まれる不要な情報の影響低減を図った．

表 1: 拡張固有表現階層（一部抜粋）

上層	中層	下層	語の例
人名	人名	人名	夏目漱石
製品名	キャラクター名	キャラクター名	ゴジラ
製品名	芸術作品名	番組名	鉄腕アトム
製品名	芸術作品名	文学名	西遊記

3.2 提案モデル

先行研究 [18] では，「文学名」と「映画名」は同時に付与されやすいといった，同一階層内のラベル間の相関（「横のつながり」）を考慮したモデルを提案している．しかし，拡張固有表現階層 [13]（表 1）のような階層的な固有表現ラベル体系において，同一階層内のラベル間の相関だけでなく，「キャラクター名」（中層）と「番組名」（下層）のように階層をまたがった固有表現ラベル間にも相関がある（「縦・斜めのつながり」がある）と考えられる．そこで本研究では，同一階層内だけではなく階層間にも相関があるとの仮定の下，マルチタスク学習を適用したモデルを提案する（図 1a）．提案モデルでは，ラベル推定に共通の隠れ層を用いることで，上層・中層・下層の固有表現ラベルの推定に際して他階層から学習された情報も共有できる．

また，階層間の相関を学習に導入した際の効果を比較検証するために，上層・中層・下層の固有表現ラベルを独立に推定するモデルを構築した（図 1b）．このモデルは各階層の推定にそれぞれ別々の隠れ層を用いているため，階層間の学習パラメータは共有されない．

出力層は，固有表現ラベルとラベル付与数を同時に推定する構造とした [16]．

なお，本研究では，階層構造を持つ固有表現ラベル体系として，拡張固有表現階層 [13] を用いた（表 1）．拡張固有

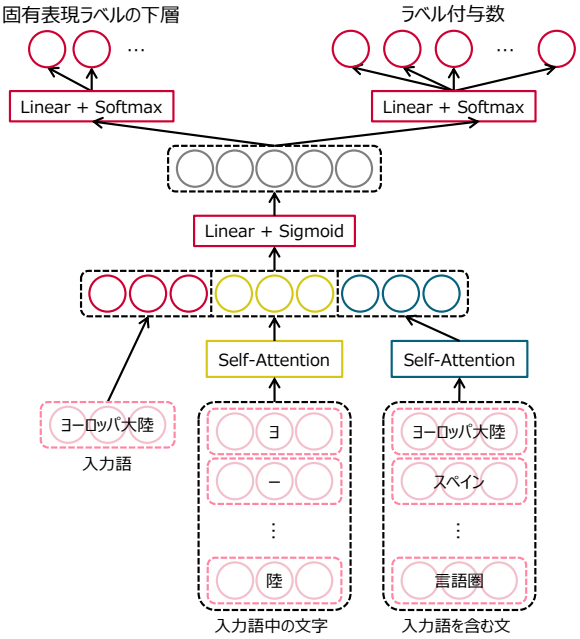


図 3: 特徴量の比較実験に用いたモデル

表現階層は，情報抽出や質問応答など幅広い自然言語処理分野への応用を目的として構築された，3 階層から成る固有表現ラベルの階層である．各階層はそれぞれ上層 28 種類・中層 103 種類・下層 200 種類の固有表現ラベルを持つ．また，拡張固有表現階層の長所として，細粒度（下層 200 種類）である点，階層化されており粒度を調整できる点が挙げられ，固有表現ラベル体系として実利用に適している．

4. 実験

4.1 データ

特徴量及び提案モデルの有効性を検証するため，先行研究 [5][16][18] に倣い，Wikipedia の記事名に対して，固有表現ラベルを推定する実験を行った．教師データとして，「拡張固有表現＋ Wikipedia」[19] を利用した．本データは，拡張固有表現階層に “CONCEPT” と “IGNORED” を加えた固有表現ラベルを，Wikipedia 記事名約 2 万項目に人手で付与したデータセットである．ラベル推定にあたって，元の拡張固有表現階層には存在しない “CONCEPT” と “IGNORED” に対しては，上層・中層ともにそれぞれ “CONCEPT” と “IGNORED” を割り当てた．

4.2 設定

まず，特徴量の変更がどの程度有効であるかを評価するために，「推定対象語のみ」,「推定対象語+文情報」,「推定対象語+文字情報」,「推定対象語+文情報+文字情報」を入力特徴量とした 4 パターンにおいて，先行研究 [16] と同様のアーキテクチャ（図 3）で固有表現ラベル推定実験を実施した，隠れ層の次元数は 200 次元とし，語の分散表現は固定した．

表 2: 入力特徴量の比較結果

	Precision	Recall	F1
$F_v + F_s + F_c$	**.9115**	**.9064**	**.9089**
$F_v + F_s$	.9110	.9060	.9085
$F_v + F_c$	.9074	.9025	.9049
F_v	.9075	.9026	.9051
水木・榊 [16]	.8943	.8920	.8907

次に，固有表現階層間の相関が推定性能に与える影響を評価するために，図 1a と図 1b に示したモデルを用いた固有表現ラベル推定実験を実施した．学習時には，上層・中層・下層の推定結果に対する損失を計算し，評価時には下層（202 種類）の推定結果のみを評価した．

計算時間の都合から，両実験とも，入力文は Wikipedia から無作為に抽出した 1000 万文を用い，学習は 1 エポックとした．また，適合率，再現率，F1 スコアを評価指標として用いた．実装には PyTorch[11] を利用した．

4.3 結果と考察

特徴量の違いによる結果を表 2 に示す．なお，特徴量 F_v を推定対象語，特徴量 F_s を文，特徴量 F_c を語の文字ユニグラムの分散表現とする．提案特徴量を入力した場合の事例ベースでの結果が，いずれも先行研究を上回った．F_v のみでも性能が向上したのは，3.1 で述べた分散表現の作成方法や次元数が影響していると考えられる．また，推定対象語 F_v に対して，語の文字ユニグラム F_c を組み合わせるだけだと分類性能が下がるが，文 F_s も組み合わせることで，わずかではあるが性能向上が見られた．

図 4 に示した例は，語「ヨーロッパ大陸」の固有表現ラベル（「大陸名」）を推定したときの，語の文字ユニグラム（図 4a）と文（図 4b）に対する Self-Attention の結果である．入力文は，「これはヨーロッパ大陸のスペインを発祥とする言語であるが、17 世紀のスペインによる新大陸の植民地化を経て、南アメリカおよび北アメリカ南部における広大な言語圏を獲得した。」とし，3.1 で述べた分割を行った．色が濃い語・文字ほど，推定の際に強く注意が向いていることを表している．語の文字ユニグラム（図 4a）について，「ヨーロッパ大陸」という語の推定に際して，語の最後の名詞「陸」に強い注意が向いているが，Higashinaka らの研究 [5] でも，固有表現ラベル推定において「語の最後の 1 文字」が特徴量として有効であることが示されており，この効果を Self-Attention により自動獲得したと考えられる．データセット中にない語に対してラベル推定を実施した際にも，「宜野湾市」（市区町村名）のように，語自身にラベルの手掛かりがある語は正しく推定できたことから，文字情報を利用することの有用性を定性的に確認できた．文（図 4b）について，「植民地化」や「言語圏」といった語に強く注意が向いている一方，「発祥」や「広大」と

表 3: 提案モデルの比較結果

	Precision	Recall	F1
3 階層を同時に推定（図 1a）	**.9130**	**.9079**	**.9104**
3 階層を別々に推定（図 1b）	.9076	.9027	.9051

表 4: 誤ったラベルの推定例

推定対象語	正解ラベル	推定ラベル
目	動物部位名	CONCEPT
池袋	国内地域名	市区町村名
ど根性ガエル	番組名・文学名	番組名

いった語への注意は弱く，Self-Attention 導入の効果が見られる．

提案モデルによる推定結果を表 3 に示す．学習に使用する特徴量は，上記比較実験の結果を踏まえ，$F_v + F_s + F_c$ を用いた．事例ベースの結果において，3 階層をそれぞれ推定するモデルよりも，まとめて推定するモデルの方が評価指標が高かった．また，下層のみを推定するモデルの結果（表 2）よりも．評価指標が高かった．この結果から，同一階層内の情報だけでなく，階層間の情報も，固有表現ラベル推定タスクにおいて有効であることが示された．

入力特徴量を $F_v + F_s + F_c$ としたときの，3 階層を同時に推定するモデル（図 1a）の学習曲線を図 5 に示す．学習は早期に収束するが，Precision/Recall 共に，10 エポック目まではわずかながらラベル推定精度の向上が見られた．本研究では学習を 1 エポックとしたが，学習回数を重ねる効果はあると考えらえる．

誤ってラベルが推定された結果の一部を表 4 に示す．普通名詞に対して “CONCEPT” が誤って推定されるケースや，類似した意味の固有表現ラベルが推定されてしまうケースが見られた．本研究では，教師データ宙における固有表現ラベルの出現頻度や，ラベルが属する階層を考慮しなかったが，より正確に固有表現ラベルを推定するためには，推定されにくい固有表現ラベルに重みを付ける，上層・中層・下層のラベルを区別するなど，各階層・各ラベルの重みを調整する必要があると考えられる．また，複数のラベルが付与されるべき語に対して，ひとつのラベルしか付与されないことがあった．これは，ラベル付与数に対する損失関数を過大評価することで改善すると考えられる．

5. まとめ

本研究では，語に対する固有表現ラベル推定にあたり，文情報及び文字情報を特徴量として利用した．また，拡張固有表現階層の階層内・階層間の相関関係を考慮したモデルの提案を行った．さらに，Wikiepdia 記事名の分類タスクの実験を行い，提案特徴量・提案モデル共に，ラベル推定に有効であることを示した．

Wikipedia に拠らない特徴量として，本研究では，文情報や推定対象語の文字ユニグラムを利用したが，その他にも，品詞情報，日本語語彙大系 [17] に基づいた意味体系，既存の固有表現抽出器により推定された IREX ラベルといった特徴量の有用性が示されている [5]．今後は，これらの特徴量も加えた評価を実施予定である．

参考文献

[1] Caruana, R.: Multitask learning, *Machine learning*, Vol. 28, No. 1, pp. 41–75 (1997).
[2] Dakka, W. and Cucerzan, S.: Augmenting wikipedia with named entity tags, *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I* (2008).
[3] Fuchi, T. and Takagi, S.: Japanese morphological analyzer using word co-occurrence: JTAG, *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 409–413 (1998).
[4] Grishman, R. and Sundheim, B.: Message understanding conference-6: A brief history, *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Vol. 1 (1996).
[5] Higashinaka, R., Sadamitsu, K., Saito, K., Makino, T. and Matsuo, Y.: Creating an extended named entity dictionary from Wikipedia, *Proceedings of COLING 2012*, pp. 1163–1178 (2012).
[6] Lee, C., Hwang, Y.-G., Oh, H.-J., Lim, S., Heo, J., Lee, C.-H., Kim, H.-J., Wang, J.-H. and Jang, M.-G.: Fine-grained named entity recognition using conditional random fields for question answering, *Asia Information Retrieval Symposium*, Springer, pp. 581–587 (2006).
[7] Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B. and Bengio, Y.: A Structured Self-attentive Sentence Embedding, *CoRR*, Vol. abs/1703.03130 (online), available from ⟨http://arxiv.org/abs/1703.03130⟩ (2017).
[8] Ling, X. and Weld, D. S.: Fine-Grained Entity Recognition. (2012).
[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013).
[10] Nothman, J., Curran, J. R. and Murphy, T.: Transforming Wikipedia into named entity training data, *Proceedings of the Australasian Language Technology Association Workshop 2008*, pp. 124–132 (2008).
[11] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A.: Automatic differentiation in PyTorch (2017).
[12] Sekine, S. and Isahara, H.: IREX: IR & IE Evaluation Project in Japanese., Citeseer (2000).
[13] Sekine, S. and Nobata, C.: Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy., *LREC*, Lisbon, Portugal, pp. 1977–1980 (2004).
[14] Toral, A. and Munoz, R.: A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia, *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources* (2006).
[15] 杉原大悟，増市博，梅基宏，鷹合基行：Wikipedia カテゴリ階層構造の固有名詞分類実験における効果，研究報告情報学基礎（FI），Vol. 2009, No. 2, pp. 57–64（オンラ

ヨーロッパ大陸

(a) 語の文字ユニグラム

ヨーロッパ大陸 スペイン 発症 言語 17世紀 スペイン 新大陸 植民地化 南アメリカ 北アメリカ 広大 言語圏

(b) 分割後の入力文

図 4: Self-Attention の結果例

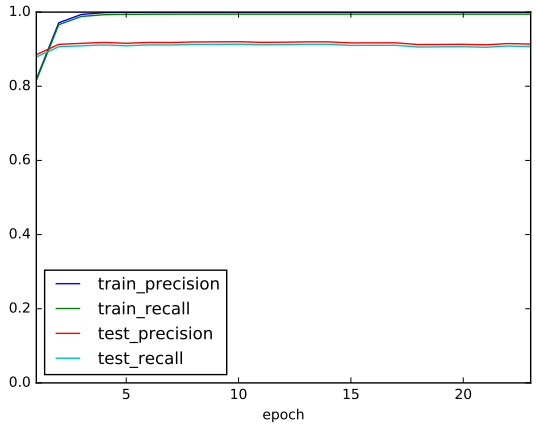


図 5: 学習曲線

イン)，入手先 ⟨https://ci.nii.ac.jp/naid/110007123991/⟩ (2009).
[16] 水木栄，榊剛史：汎用性を志向した Wikipedia エントリへの拡張固有表現付与 (思考と言語)，電子情報通信学会技術研究報告 = IEICE technical report : 信学技報， Vol. 117, No. 81, pp. 47–52（オンライン)，入手先 ⟨https://ci.nii.ac.jp/naid/40021249243/⟩ (2017).
[17] 悟池原，NTT コミュニケーション科学研基礎究所：日本語語彙大系，Iwanami EP CD-ROM, 岩波書店 (1999).
[18] 鈴木正敏，松田耕史，関根聡，岡崎直観，乾健太郎：Wikipedia 記事に対する拡張固有表現ラベルの多重付与，言語処理学会第 22 回年次大会，pp. 797–800 (2016).
[19] 関根聡，安藤まや，松田耕史，鈴木正敏，乾健太郎：「拡張固有表表現＋ Wikipedia」データ，言語処理学会第 22 回年次大会，pp. 41–44 (2016).