深層学習を用いた実世界参照による分野特有の固有表現の認識

友利 涼† 二宮 崇†† 森信介†††
†愛媛大学 工学部 情報工学科
††愛媛大学 大学院理工学研究科 電子情報工学専攻
†††京都大学 学術情報メディアセンター
{tomori@ai., ninomiya@}cs.ehime-u.ac.jp, forest@i.kyoto-u.ac.jp

概 要

本論文では、将棋の解説に対する固有表現を題材として、テキスト情報に加えて実世界情報を参照する固有表現認識を提案する。この題材での実世界情報は、固有表現認識の対象となる解説が言及している将棋の局面である。局面は、盤面上の駒の配置と持ち駒であり、すべての可能な状態がこれによって記述できる。提案手法では、まず各局面の情報をディープニューラルネットワークの学習方法の 1 つである stacked-auto-encoder を用いて事前学習を行う。次に、事前学習の結果をテキスト情報と組み合わせて固有表現を認識する。提案手法を評価するために、条件付き確率場による方法等との比較実験を行った。実験の結果、提案手法は他の手法よりも高い精度を示し、実世界の情報を用いることにより固有表現認識の精度向上が可能であることが示された。提案手法は、ニュース映像に対する一般の固有表現認識や検査データが付随するカルテに対する医療固有表現認識等に応用することが可能である。

1 はじめに

近年、情報化技術の発展により、インターネット上やデータベース上にはテキストとそのテキストに付随する実世界情報が大量に存在している。シンボルグラウンディング問題 [1] のように、自然言語処理に表層的なテキスト処理のみでは限界があると考えられており、実世界の情報をいかに用いるかが課題となっている。実世界の情報を用いる研究として画像とテキストを結び付けること [2, 3, 4] などが行われている。

本論文では、実世界の情報を参照することにより自然言語処理の精度を向上させることが可能であることを示す。具体的な自然言語処理としては、固有表現の認識を課題とする。固有表現とは、新聞などの検索を目的とする人名や地名など約 8 種類の定義 [5] が一般的であるが、近年では医療の言語処理などを目的としたバイオ固有表現 [6] なども提案されている。

本論文で提案する実世界情報の参照の効果の確認のためには、人手で固有表現タグが付与されたテキストに実世界の情報が付随するコーパスが必要である。しかしながら、このようなコーパスはほとんどなく、本論文では、この条件を満たす将棋解説コーパスを題材とする。この題材での固有表現は、後述するように戦型などの 21 種類であり、実世界情報は将棋の局面である。

実世界情報を参照する固有表現認識器の実現のために、ディープニューラルネットワークの枠組みを用いる。まず実世界情報だけを用いて stacked-auto-encoder と呼ばれるディープニューラルネットワークのための事前学習を行う。続いて、これをテキスト情報のみを参照する通常の固有表現認識器と統合し、実世界も参照するニューラルネットワークを構築する。最後に、統合したニューラルネットワークの再学習を行う。なお、本論文で提案する実世界を参照する固有表現認識器は、固有表現一般に適用することが可能である。

2 関連研究

Kameko らは、将棋局面情報に将棋解説文が付与された将棋解説コーパスを用いて、深層学習を応用した将棋解説文のための単語分割手法を提案している [7]。この手法は、まず将棋局面情報を入力とする深層学習を行うことで将棋用語辞書を獲得し、得られた将棋用語辞書を用いて単語分割を行うことで単語分割の精度向上を実現している。具体的には、将棋用語辞書を作

種類	意味
Hu	人 (対局者や解説者を含む人)
Tu	手番 (例: 先手, 後手, ▲, △)
Po	位置 (81 通りと「駒台」「同」など)
Pi	駒 (成り駒を含め有限)
Ps	駒の指定 (例: 右, 直)
Mc	動きの明確化 (例: 成, 不成)
Pa	駒の属性 (例: 道, 利き, 頭)
Pq	駒の数 (例: 1 枚, 切れ)
Re	盤面の領域 (例: 中央, 駒台, 4 筋, 3 段 目)
Ph	対局の進行 (例: 序盤, 中盤, 終盤)
St	戦型 (「棋士名＋流」も含む)
Ca	囲い (例: 矢倉, 美濃 囲い)
Me	指し手評価
Mn	指し手別名
Ee	評価要素 (部分の評価のみ)
Ev	形勢評価 (局面全体の判断のみ)
Ti	時間 (概数表現を含む)
Ac	対局者が主語の述語
Ap	駒が主語の述語
Ao	その他の表現が主語の述語
Ot	その他

表 1: 将棋の固有表現の種類とその意味

成するために、まず確率的単語分割手法により確率的に単語分割されたコーパスを生成し、次に、将棋局面情報を入力とし、確率的コーパスの単語を出力とするディープニューラルネットワークを学習し、学習されたディープニューラルネットワークを用いて単語候補をスコア付けし、抽出することで、将棋用語辞書が得られる。本研究と彼らの研究との違いは次の通りである。(1) 彼らは将棋の局面との対応を参照して語彙を獲得しておくことで単語分割しているが、本研究では将棋の局面との対応を参照して固有表現解析を行うこと、(2) 彼らは各文の単語分割には局面を直接参照していないが、本研究では固有名解析に局面を直接参照すること、(3) 本論文では、動的に変化する実世界を解析時に参照していることが挙げられる。

3 将棋解説コーパス

将棋は 2 人で行うボードゲームで 9×9 のマスの盤面と成った駒も含めて 14 種類の駒を用いる。盤面上の駒の配置と持ち駒 (局面と呼ぶ) からゲームの状態に関するすべての情報が得られる完全情報ゲームである。将棋にはプロ制度があり、日々多数のプロ間の対局が行われている。多くの対局には、対局者以外のプロが解説を行い、その解説文がインターネットで配信されている。

本論文で用いるコーパスは、将棋の解説文に対して単語分割と固有表現タグ付与を人手で行ったものである。固有表現は、将棋解説に特化されており、表 1 のように 21 種類が定義されている。実際のアノテーションは、BIO2 形式であり、各単語ごとに BIO タグ [1] が 1 つ付与されている。

各解説文には、対象とする局面が対応している。局面の情報は、解説が言及する実世界の情報とみなすことができ、これを参照することによる固有表現認識の精度向上が本論文の中心となるアイデアである。

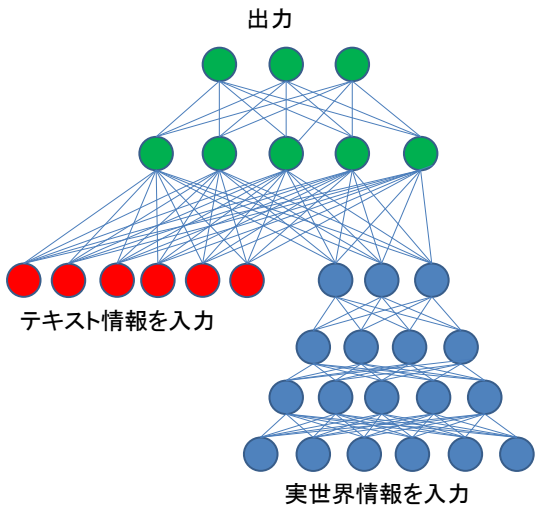

図 1: 実世界を参照する固有表現認識のためのディープニューラルネットワーク

4 提案手法

将棋解説文の固有表現認識を行うために用いたニューラルネットワークと実世界 (将棋の局面) の事前学習について説明する。

4.1 ディープニューラルネットワークの構成

図 1 は実世界を参照する固有表現認識のニューラルネットワークの全体図である。図の右下の 4 つの層は実世界に関するニューラルネットワークである。図の上側のニューラルネットワークでテキスト情報と実世界情報を統合して固有表現認識を行う。実世界に関するニューラルネットワークの中間層の層数については開発データを用いて調整する [2]。

[1] B (Begin) はある固有表現の最初の単語、I (Intermediate) は同種の固有表現の継続、O (Other) はいずれの固有表現でもないことを意味する (合計: 21 × 2 + 1 = 43 種類)。

[2] 後述する実験では、層数を 1 層から 4 層まで変えて実験した結果、中間層 3 層の場合が最も精度が高かったため、テストデータに対する実験では中間層 3 層を用いて評価した。

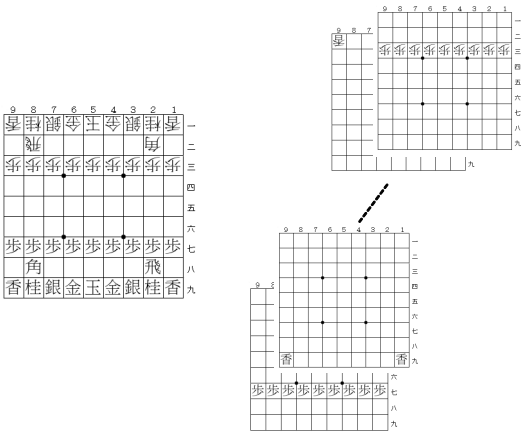

図 2: 将棋盤面の素性

素性
w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}
$w_{i-2}w_{i-1}$, $w_{i-1}w_i$, w_iw_{i+1}, $w_{i+1}w_{i+2}$
$w_{i-2}w_{i-1}w_i$, $w_iw_{i+1}w_{i+2}$
$type(w_{i-2})$, $type(w_{i-1})$, $type(w_i)$,
$type(w_{i+1})$, $type(w_{i+2})$
$pos(w_{i-2})$, $pos(w_{i-1})$, $pos(w_i)$,
$pos(w_{i+1})$, $pos(w_{i+2})$

表 2: タグを推定している w_i のテキスト素性

4.2 実世界情報の事前学習の入力とテキストの入力

実世界情報は、図 1 の右下に示されるニューラルネットワークの入力として参照される。将棋の言語処理では、図 2 に示すように盤面上の全てのマス (9×9) における先手と後手を区別した駒の種類 (2×14) の有無に対応する $2,268 = 9 \times 9 \times 2 \times 14$ 次元のバイナリ素性と、持ち駒を記述する先手と後手の 7 種類の駒の個数に対応する 14 次元の整数素性 ($14 = 7 \times 2$) とする。

テキストの素性には 1-of-k 表現を用いた。表 2 に実験で用いたテキストの素性を示す。w_i は現在タグを推定している単語である。$type(w)$ は w の文字の種類、ひらがな、カタカナ、漢字、数字、記号を表し、$pos(w)$ の w の品詞を表す。

4.3 実世界情報の事前学習

本研究では事前学習のひとつである stacked-auto-encoder を用いて実世界情報の事前学習を行った。図 3 の左側のように、まず 3 層のニューラルネットワークの入力層と出力層に同じ実世界情報を与え、入力と同じ出力を予測するニューラルネットワークを学習する。このとき中間層の次元数 (ノード数) を入力層より

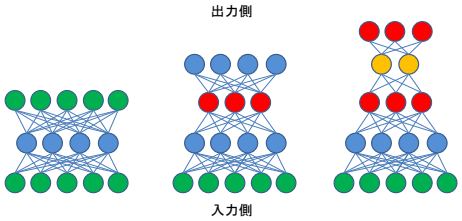

図 3: Stacked-auto-encoder.

も少ない次元数とすることで、中間層において次元圧縮された実世界情報を得ることができる。次に図 3 の中央のように新しい中間層を 1 つ増やし、先に得られた中間層から同じ中間層を予測するニューラルネットワークを構築する。このとき、先に得られた中間層に対する再学習は行わず、新しい中間層に対する学習のみ行う。同様に、層を積み重ねることでより、より深く一般化された特徴を学習することができる。本研究では、実世界情報のみを用いて事前学習を行なうこととする。

4.4 テキスト情報と実世界情報の統合

事前学習が終わると、テキスト情報側のニューラルネットワークと実世界情報側のニューラルネットワークを統合し、固有表現認識タスク用にファインチューニングを行う。ファインチューニングにおいては、各単語の正解の BIO タグの尤度が最大になるように、図 1 に示されているニューラルネットワーク全体が学習される。したがって、固有表現認識のための中間層だけでなく、事前学習された実世界情報側のニューラルネットワークも同時に調整される。

4.5 最適タグ列の推定

固有表現認識は各単語に対する BIO タグを推定することで実現されるが、単語ごとに BIO タグを推定すると、推定した固有表現が I タグから始まってしまう場合など、BIO タグ制約を満たさない場合がある。本研究では BIO タグを満たすために、各タグの推定確率を用いて、BIO タグの制約を満たす遷移のみを許すビタビアルゴリズムを適用し [8]、BIO タグ列を推定する。

5 評価実験

提案手法の有効性を確認するために、固有表現認識の実験を行った。この節では従来手法と提案手法を説

	文数	単語数	固有表現数
学習	1,546	27,025	7,922
テスト	492	7,161	2,365

表 3: 将棋解説コーパスの諸元

層	0	1	2	3	4
次元数	2,282	1,000	500	200	100

表 4: 将棋盤面埋め込みに用いたニューラルネットワークの各層の次元数

明し、それぞれの結果を提示する。実験には将棋解説コーパスを用いた (3 節参照)。表 3 に実験で用いたコーパスの詳細を示す。ニューラルネットワークの実装には Chainer[3] を用いた。表 4 に局面を事前学習する際に用いた各層の次元数を示す。また、既存手法として条件付き確率場 (CRF) [9] による方法を評価し、提案手法と比較した。CRF による手法の実装には CRF++[4] を用いた。各単語の品詞推定には KyTea[5] を用いた。

表 5 に実験の結果を示す。“DNN” はディープニューラルネットワークを、“V” は BIO タグの遷移制限付きビタビアルゴリズム (4.5 節参照) を、“B” は局面の参照 (4.3,4.4 節参照) を示す。表 5 より、ビタビアルゴリズムを適用するしないに関わらず CRF よりも DNN の方が精度が高いことがわかる。テキスト情報と実世界情報を用いた提案手法が、テキスト情報のみを用いた従来手法よりも精度が高いことが確認できる。実世界情報を用いることで、“CRF” や “DNN” よりも “DNN + B” の方がそれぞれ 0.87Pt、0.35Pt 高く、“CRF + V” や “DNN + V” よりも “DNN + V + B” の方がそれぞれ 1.27Pt、1.16Pt 高くなった。

6 おわりに

本論文では、新たな固有表現認識の解法として、実世界情報の参照を提案した。提案手法では、全体の枠組みとしてディープニューラルネットワークを用いる。まず実世界情報だけを用いて stacked-auto-encoder の事前学習を行い、これをテキスト情報のみを参照する通常の固有表現認識器と統合し、統合したニューラルネットワークの再学習を行う。

実験では、各文に駒の配置という実世界情報が付随し、さらに固有表現タグが付与された将棋解説コーパスを用いた。提案手法を実装し、既存手法に対する優位性や実世界情報参照の効果を実験的に示した。

本論文で提案した実世界情報を参照する固有表現認識の手法は、実世界情報のためのネットワークを変更することで一般の固有表現認識にも適用することが可能である。

[3] http://chainer.org/ (2016/01/06 アクセス).

[4] https://taku910.github.io/crfpp/ (2016/01/06 アクセス).

[5] http://www.phontron.com/kytea/ (2016/01/06 アクセス).

手法	BIO タグ推定精度	適合率	再現率	F 値
CRF	90.36%	90.37%	78.13%	83.75
CRF + V	90.36%	90.14%	78.52%	83.93
DNN	90.81%	89.77%	79.40%	84.27
DNN + V	90.81%	91.61%	77.63%	84.04
DNN + B (提案手法)	91.04%	89.72%	80.08%	84.62
DNN + V + B (提案手法)	91.04%	88.88%	81.81%	85.20

表 5: 固有表現認識の結果

謝辞

本研究は JSPS 科研費 25280084, 26540190 の助成を受けたものである。ここに謝意を表する

参考文献

[1] Harnad, S.: The Symbol Grounding Problem, *Physica D*, Vol. 42, pp. 335–346 (1990).

[2] Karpathy, A., Joulin, A. and Li, F. F. F.: Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, *Advances in Neural Information Processing Systems 27*, pp. 1889–1897 (2014).

[3] Socher, R., Karpathy, A., Le, Q. V., Manning, C. D. and Ng, A. Y.: Grounded Compositional Semantics for Finding and Describing Images with Sentences, *TACL*, Vol. 2, pp. 207–218 (2014).

[4] Kennington, C. and Schlangen, D.: Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution, *Proc. of ACL '15*, pp. 292–301 (2015).

[5] Sang, E. F. T. K. and Meulder, F. D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. of the CoNLL2003*, pp. 142–147 (2003).

[6] Settles, B.: Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets, *Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 33–38 (2004).

[7] Kameko, H., Mori, S. and Tsuruoka, Y.: Can Symbol Grounding Improve Low-Level NLP? Word Segmentation as a Case Study, *Proc. of EMNLP 2015*, pp. 2298–2303 (2015).

[8] Sasada, T., Mori, S., Kawahara, T. and Yamakata, Y.: Named Entity Recognizer Trainable from Partially Annotated Data, *Proc. of PACLING 2015* (2015).

[9] Lafferty, J. D., McCallum, A. and Pereira, F. C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. of ICML '01*, pp. 282–289 (2001).