実世界情報を参照した分野特有の固有表現体系の自動獲得

友利 涼† 森 信介††
† 京都大学大学院 情報学研究科
†† 京都大学 学術情報メディアセンター
tomori.suzushi.72e@st.kyoto-u.ac.jp, forest@i.kyoto-u.ac.jp

1 はじめに

自然言語処理の多くのタスクにおいて、機械学習技術を用いる手法が大きな成功を収めており、アノテーション付きコーパスの重要性が高まっている。しかし、アノテーション付きコーパスが十分に存在する分野は限られており、新しくコーパスを作成することはコストが大きく、その分野の専門知識が求められる。

また、近年、情報技術の発展により、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。

そこで本研究では、実世界情報を参照することで、分野特有の専門知識を用いずに固有表現体系の獲得を目指す。本論文では、将棋解説文とその解説文が参照する局面を実世界情報として用いて、単語列をクラスタリングすることで、分かち書きされていない生テキストから固有表現体系を抽出する。

2 関連研究

固有表現の自動獲得の手法には、ブートストラップ [2] や Co-Training [3] がある。ブートストラップは、与えられた固有表現のインスタンス (用語) と共起するパターンを抽出し、新たなインスタンスとして追加する。Co-Training は複数の分類器を異なる素性から作成し、各分類器が出力した確信度の高い固有表現を互いの訓練データに追加し、分類器の再学習を行う。これらの手法は半教師あり学習であり、少量のアノテーションされたデータから大量の用語を獲得する手法である。

テキストと実世界情報が紐付いたデータをクラスタリングする手法もいくつか提案されている。Latent semantic analysis を拡張した手法 [4] やスペクトラルクラスタリングを拡張した手法 [5]、Matrix factorization を用いる手法 [6] が提案されている。

将棋解説文と局面を用いた自然言語処理には、単語分割 [7] や固有表現認識 [8] がある。将棋解説文の単語分割では、ニューラルネットワークを用いて単語候補と局面を対応づけすることにより、将棋専用の辞書を作成する。作成した辞書を用いることで将棋解説文の単語分割の精度が向上した。将棋解説文の固有表現認識の研究では、固有表現を推定する際にテキスト情報だけではなく、解説文が参照する局面をニューラルネットワークへの入力の一部として用いることで精度が向上することを示した。

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。

本論文で用いる将棋解説コーパス [9] は、将棋の解説文とその解説文が参照する局面が対応付けされており、局面の情報は、解説文が言及する実世界の情報とみなすことができる。また、一部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。固有表現は人名や戦型名、囲い名など 21 種類が定義されている。固有表現が付与された単語列 (用語) は将棋解説文の特有の概念を持っており、複数の単語から構成される場合がある。表 1にコーパスの詳細を示す。

表 1: 将棋解説コーパスの諸元

	文数	単語数	固有表現数 (異なり数)	局面数
全テキスト (アノテーションなし)	667,362	-	-	273,303
アノテーションありテキスト	2,038	34,186	1,799	547

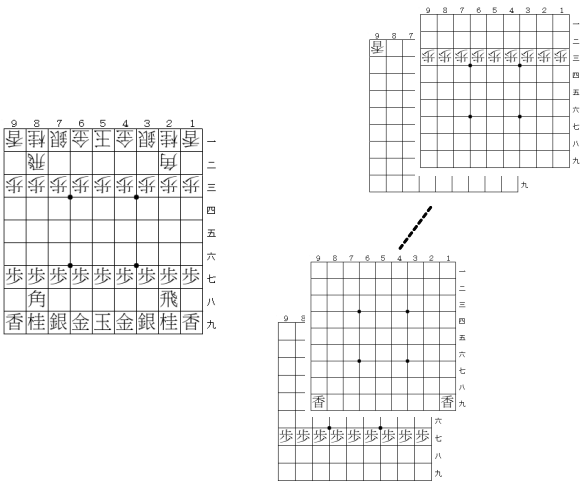

図 1: 将棋局面の素性

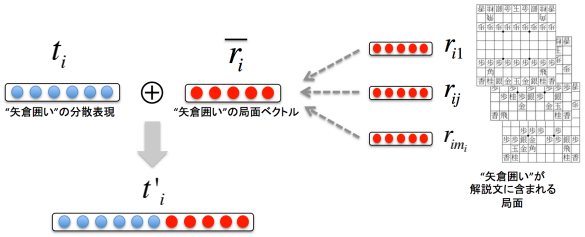

図 2: 提案手法の概要

4 提案手法

本研究では実世界情報を参照したクラスタリングを行う。本研究では将棋解説文のデータを用いており、本節では将棋局面から素性を作成する方法と提案手法を述べる。

4.1 将棋局面素性

将棋局面から局面の素性作成の方法を図 1に示す。局面の素性 r_{ij} は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。

4.2 用語クラスタリング

本節では実世界情報を参照した用語クラスタリングについて説明する。用語のクラスタリングとは複数単語からなる用語を分類することである。提案手法の概要を図 2に示す。提案手法では用語クラスタリングは以下の手順に従って行われる。

1. 生テキストから用語候補を取得する
2. 用語候補の分散表現 t_i を取得する
3. 用語候補の分散表現 t_i とその用語候補の実世界情報 (局面ベクトル) $\overline{r_i}$ を組み合わせ、t'_i を作成する
4. t'_i $(i = 1, ..., V)$ をクラスタリングする

用語は複数単語列から構成されるため、名詞間の連接情報を用いて用語候補を抽出する [10]。まず、分かち書きされていないコーパスを形態素解析して単語に分割し、単語列とその品詞情報を取得する。その後、名詞の連続を複合名詞として抽出し、以下に示す重要度に基づき用語候補を抽出する。

ある単語 w の直前に付き、w と複合名詞を作る異なる単語の数を $Pre(w)$ とし、w の直後に付き、w と複合名詞を作る異なる単語の数を $Post(w)$ とする。ある複合名詞 $w_1w_2w_3...w_n$ の重要度 Imp は式 (1) で与えられる。

$$Imp(w_1w_2w_3...w_n) = \left[\prod_{i=1}^{n}\{(Pre(w_i)+1)\cdot(Post(w_i)+1)\}\right]^{\frac{1}{2n}} \quad (1)$$

$Pre(w)$ と $Post(w)$ に $+1$ しているのは、$Pre(w)$ または $Post(w)$ が 1 つでも 0 になると、Imp の値が 0 になることを避けるためである。この Imp とその複合名詞の出現頻度の積をスコアとし、上位 10%を用語候補として抽出する。

用語の分散表現の獲得には word2vec [11, 12] と同様のアルゴリズムを用いた。類似した用語は類似した周辺単語を持っていると仮定し、用語候補の周辺に出現

表 2: 用語抽出精度

	適合率	再現率	F 値
用語抽出	0.427	0.325	0.369

する単語を予測するニューラルネットワークを用いて用語候補の分散表現を取得する。

用語候補の分散表現 t_i とその用語候補の局面ベクトル $\overline{r_i}$ を連結し t'_i を得る。ある用語候補 t_i に対応する $\overline{r_i}$ は、その用語候補が出現する将棋解説文が参照している将棋局面を4.1節に示した方法で作成された素性ベクトルの平均から与えられる。

例えば、「矢倉囲い」が用語候補として抽出されたとする。まず、将棋解説文に出現する「矢倉囲い」の文字列を見つけ出し、周辺単語から分散表現を獲得する。また、「矢倉囲い」の文字列が出現する解説文が対応する局面を全て抽出し、それぞれの局面から素性を作成する。抽出した局面数が m 個あるとすると、「矢倉囲い」の局面ベクトルの各次元の値は、$(r_{i1}, ..., r_{ij}, ..., r_{im})$ の各次元の値の平均から計算される。

最後に t'_i $(i = 1, ..., V)$ を k-means 法を用いてクラスタリングすることで、各用語候補は 1 つのクラスタに分類される。

5 実験

本節では、提案手法を評価するために行った実験について述べる。

5.1 実験設定

本実験では、3節に示した将棋解説コーパス[1] を用いて実験を行った。このコーパスで定義されている固有表現を正解クラスとして評価した。しかし、将棋解説コーパスでは同じ用語でも文脈により複数の固有表現になりうる。今回の実験では簡単化のために、ある用語が複数の固有表現に属する場合にはもっとも出現頻度が高かった固有表現を正解クラスとした。また、ある固有表現に属する用語がコーパス中で 10 種類以下の場合、その固有表現はクラスタリングの評価の対象外とし、実験では 16 種類の固有表現を正解クラスとした。

本実験では、まず生テキストから4.2節で述べた方法で用語候補を抽出した。形態素解析には KyTea[2] を用いた。その後、抽出した用語候補の分散表現を将棋解説コーパスの全テキストを用いて計算した。このとき、テキストは分かち書きされていないため、KyTea を用いて単語分割し、用語候補の分散表現の次元数は 100、前後 3 単語を予測するニューラルネットワークを用いた。次に局面を参照したベクトルを連結し、k-means 法でクラスタリングを行った。k-means 法は分類するクラスタの数をハイパーパラメータとして与える必要がある。本実験では、正解クラスタの数に合わせて 16 とした。比較手法としてテキスト情報のみのベクトル t_i $(i = 1, ..., V)$ をクラスタリングした。提案手法と同様に用語候補の分散表現の次元は 100、16 クラスタに分類した。クラスタリングの評価は、生テキストから抽出した用語候補の内、将棋解説コーパスに人手でアノテーションされた用語と一致する単語列について行う。

5.2 評価尺度

用語抽出の評価尺度には以下の式で表される再現率と適合率、F 値を用いた。将棋解説コーパスのアノテーション付きテキストに出現した固有表現を正解として評価した。

$$\begin{aligned} \text{再現率} &= \frac{\text{正解数}}{\text{用語の総数}} \\ \text{適合率} &= \frac{\text{正解数}}{\text{用語候補の総数}} \\ \text{F 値} &= \frac{2 \cdot \text{再現率} \cdot \text{適合率}}{\text{再現率} + \text{適合率}} \end{aligned}$$

ただし、正解数とは用語候補の内、将棋解説コーパスに出現した用語と一致する数である。

クラスタリングの評価には以下の式で表される Entropy と NMI、Purity、Inverse purity、F-measure を用いた。いずれの評価尺度も 0 から 1 までの値をとり、NMI と Purity、Inverse purity、F-measure は値が大きいほど結果がよく、Entropy は値が小さいほど結果が良い。

$$\begin{aligned} \text{Entropy}(\omega) &= -\sum_k \frac{|\omega_k|}{|\omega|} log \frac{|\omega_k|}{|\omega|} \\ \text{NMI}(\omega, c) &= \frac{\text{相互情報量 } (\omega, c)}{\frac{1}{2}(\text{Entropy}(\omega) + \text{Entropy}(c))} \\ \text{Purity}(\omega, c) &= \frac{\sum_k \max_j |\omega_k \cap c_j|}{\sum_k |\omega_k|} \\ \text{Inverse purity}(\omega, c) &= \text{Purity}(c, \omega) \\ \text{F-measure} &= \frac{2 \cdot \text{Purity} \cdot \text{Inverse purity}}{\text{Purity} + \text{Inverse purity}} \end{aligned}$$

ただし、$\omega = \{\omega_k\}$ はクラスタリングの結果であり、ω_k は 1 つのクラスタを表す。$c = \{c_j\}$ は正解クラスであり、c_j は 1 つのクラスを表す。

[1] http://www.ar.media.kyoto-u.ac.jp/data/game/
[2] http://www.phontron.com/kytea/

表 3: クラスタリングの結果

Method	Entropy	NMI	Purity	Inverse purity	F-measure
ベースライン	0.354	0.428	0.413	0.659	0.508
提案手法	**0.352**	**0.431**	**0.417**	**0.660**	**0.512**

5.3 結果と分析

表 2の用語候補抽出の精度を示す。本実験では 1,399 個の用語候補を抽出でき、抽出した用語候補のうち、将棋解説コーパスに人手でアノテーションされた 1,799 個の用語に含まれるものは 572 個であった。

表3にクラスタリングの結果を示す。実験結果より、いずれの評価尺度でもベースラインより提案手法が精度が良いことより、テキスト情報のみを用いるよりも局面情報も参照した方が良いクラスタリングができている。

用語候補を抽出する際、コーパス中に低頻度で出現した用語は抽出することができなかった。例えば、「１時間２４分」のような時間に関する固有表現や「２四」のような盤面の位置に関する固有表現のような数字が含まれる用語は高頻度で出現するが、同じ文字列が出現する割合が低いので抽出することが難しい。このような用語が、コーパスに出現した 1,799 個の用語の内、354 個を占めたが、１つも用語候補として抽出できなかった。テキスト情報のみのクラスタリングでは、「立石流」や「山崎流」、「コーヤン流」が別々のクラスタに属していたが、提案手法では同じクラスタに属するようになった。

6 おわりに

本論文では、固有表現体系の自動獲得のために将棋解説文と将棋局面を用いて用語候補をクラスタリングする手法を提案した。提案手法では、分かち書きされていない生テキストから用語候補を抽出する。その後、用語候補の分散表現とその用語候補の実世界情報を計算し、それらのベクトルを連結する。最後に連結されたベクトルをクラスタリングする。実験結果より、用語候補だけでなく、その用語候補に関連した局面の情報を用いることでクラスタリングの精度が向上した。

本論文で提案した手法は、生テキストから自動で用語候補を抽出しており、用語候補はクラスタリングされているため、アノテーションをする際のコストを下げることができる。固有表現体系の自動獲得は、実世界情報の素性を変更することで他の分野にも適用することが可能である。

謝辞

本研究は JSPS 科研費 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] Thelen, M. and Riloff, E.: A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 214–221 (2002).

[3] Collins, M. and Singer, Y.: Unsupervised Models for Named Entity Classification, *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100–110 (1999).

[4] Xuanhui Wang, Jian-Tao Sun, Z. C. C. Z.: Latent semantic analysis for multiple-type interrelated data objects, *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, p. 236243 (2006).

[5] Long, B., Zhang, Z. M., Wú, X. and Yu, P. S.: Spectral Clustering for Multi-type Relational Data, *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA, ACM, pp. 585–592 (2006).

[6] Wang, F., Li, T. and Zhang, C.: Semi-Supervised Clustering via Matrix Factorization, SDM (2008).

[7] Kameko, H., Mori, S. and Tsuruoka, Y.: Can Symbol Grounding Improve Low-Level NLP? Word Segmentation as a Case Study, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 2298–2303 (2015).

[8] Tomori, S., Ninomiya, T. and Mori, S.: Domain Specific Named Entity Recognition Referring to the Real World by Deep Neural Networks, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 236–242 (2016).

[9] Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H. and Tsuruoka, Y.: A Japanese Chess Commentary Corpus, *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (2016).

[10] 中川裕志, 森辰則, 松崎知美, 川上大介: 日本語マニュアル文における名詞間の連接情報を用いたハイパーテキストのための索引の抽出, 情報処理学会論文誌, pp. 1986–1994 (1997).

[11] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *Proceedings of workshop at the International Conference on Learning Representations (ICLR 2013)* (2013).

[12] 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*, Vol. 26, pp. 3111–3119 (2013).