様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２９ 年　 ４ 月 ２５ 日現在

機関番号：１４３０１
研究種目：挑戦的萌芽研究
研究期間：2014～2016
課題番号：２６５４０１９０

研究課題名（和文）解説するコンピュータ将棋：データ分析と未来予測の言語化

研究課題名（英文）Verbalization of data analysis or future prediction

研究代表者
森　信介（Mori, Shinsuke）

京都大学・学術情報メディアセンター・教授

研究者番号：９０４５６７７３

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：与えられた盤面およびそこから先読みを行った結果得られる盤面に対して解説を生成方法を提案し自動解説を実現した。
この過程で得られる用語と局面の自動対応（シンボルグラウンディング）モジュールを用いて言語のキーワードによる局面検索が実現できることを示し、情報検索のトップ会議（ACM SIGIR 2017）に採択された。また、本研究テーマを通して作成した将棋の固有表現コーパスを LREC 2016 にて発表し、これを用いて、局面を参照する固有表現認識器を提案し、言語処理のトップ会議である ACL 2016 にて発表を行った。

研究成果の概要（英文）：We have proposed and realized a method for generating commentaries for a given game state and possible future states.
We have built a game corpus through this research and presented it in LREC 2016. With this corpus we have developed a method for searching game states by natural language keywords and made a presentation in the top conference of information retreivel, ACM SIG IR. We have also shonw that real world information improves the accuracy of a named entity recognizer. We made a presentation on this research result in one of the top conferences of natural language processing, ACL 2016.

研究分野：自然言語処理

キーワード：ゲーム　言語生成　情報検索　自動解説　シンボルグラウンディング

1. 研究開始当初の背景

センサー技術とデータ通信技術の進歩により、大規模なデータ (ビッグデータ) の時代を迎えている。これに伴い、データの可視化・言語化の技術が非常に重要になっている。例えば、英エコノミスト誌は charticle (chart + article; 可視化されたデータと解説文からなる記事; 図 1 参照) に力を入れている。データの可視化の研究は枚挙に暇がないが、本研究計画のようなデータの言語化の研究は皆無といってよい。今後、データに対する計算機による分析や思考を自然言語で表現する技術が必須になる。

2. 研究の目的

本応募課題では、2013 年の世界コンピュータ将棋選手権 4 位の「激指」に、己の分析や思考を自然言語で表現する能力を持たせることを目標とする。2013 年の電王戦では、コンピュータ将棋はプロ棋士に勝ち越し、トッププロといい勝負をする日も近い。将棋を題材とする利点は、計算機と同等以上の能力を持つ人間による解説文が大量に利用可能であることである。

本研究においては、まず、ある局面とプロの解説文から、局面とその解説に必要となる特徴的な表現の関係を獲得する手法を確立する (図 2 参照)。次に、仮名漢字変換 (Google 日本語入力など) に用いられる言語モデルを文生成を目的とするモデルに拡張する。このような言語モデルと局面から算出される特徴的な表現を用いて、計算機が右図のようなプロと同等の解説文を生成できるようにする。他の大規模データへの展開のために、学習に用いるプロによる解説の量と出力される解説の質との関係も明らかにする。

3. 研究の方法

本研究は、以下の 4 つの部分からなる。本応募課題の重要なポイントは **(2)**〜**(4)** であり、最難関は **(4)** 解説文の生成である。**(1)** はすべてに関わる準備で、新規性は乏しいものの確実に達成する必要がある。

(1) 言語解析の分野適応: 言語解析ツール (KyTea, EDA) の将棋解説文に対する精度を向上する。

研究代表者は、実テキストでの言語処理を重視し、様々な分野のテキストに対して容易に高い解析精度が実現可能な言語処理方法について研究し、その成果として形態素解析器 KyTea (単語分割・品詞推定・読み推定) や係り受け解析器 EDA を公開している。後段のために、これを将棋の解説文に適応し、高い解析精度を実現する。

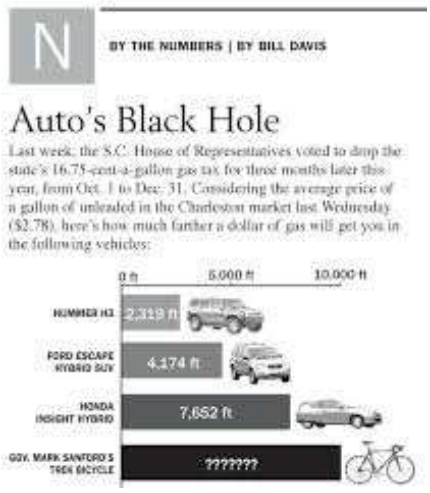


図 1: charticle の例

我々が既に一部実施した結果や、研究代表者による特許文や twitter やレシピへの適応の実績を考慮すると、大きな困難はないと想定される。

(2) 局面と特徴表現の対応獲得: プロの解説文から局面の特徴ベクトルと特徴表現の対応を獲得する。

プロ棋士による解説が付いた棋譜が一部有料ながら公開されている。2011 年以前の順位戦と名人戦に限定しても 104,903 局面あり、349,750 文の解説が得られる。このようにして得られる局面 x_i と解説文の列 $\boldsymbol{s}_i$ (各文は (1) の結果得られる言語解析器により単語列に分割) の組から、局面の次元 $\exists D$ の特徴ベクトル

$$\boldsymbol{\phi}(x) = (\phi(x)_1,\ \phi(x)_2,\ \ldots,\ \phi(x)_D)$$

と特徴表現 (単語列) $\boldsymbol{w}$ の確率的対応関係 $P_w(\boldsymbol{w}|\boldsymbol{\phi}(x))$ を獲得する。非常にシンプルな特徴量と予測モデルを用いた我々の予備実験では、戦型 (例: 中飛車, 穴熊) を中心とする特徴表現と局面との対応において、F 値として 0.38 の精度が得られている。この精度は、実用的な解説システムのための精度としては不十分であり、使用する特徴量と予測モデルを改善することによってその精度を大幅に向上させることは本研究プロジェクトの重要な課題の一つである。潜在変数を利用した生成モデルによるクラスタリングアルゴリズムや能動学習の併用することで、現実的

な人手コストで解説文に表現されている様々な概念の獲得を行う。

(3) 文生成用の言語モデルの構築: 文生成を指向し、非文の生成確率が低い言語モデルを構築する。

音声認識などに用いられる通常の言語モデルとは異なり、本応募課題の文生成モジュールの入力は、自然言語としての制約が一切ない (音声認識の入力の発音列は、自然言語としての制約を反映している)。したがって、局面から得られる特徴表現とその確信度などの少数のパラメータから文法的な文を生成する必要がある。予備的実験では、単語 n-gram モデルで最大の確率となる単語を貪欲法で選択したが、まったくもって不十分であった。

本応募課題では、文の係り受け構造を考慮した構造的言語モデルの利用を考える。この文献のままでは単語 n-gram モデルより少し文法的になる程度であると想定される。プロの解説文により近くなるように、学習データの生成確率が非常に高くなるようなモデルを研究する。

(4) 解説文の生成: 将棋プログラム「激指」による現局面などに対する先読みから解説文を生成する。

計画 (2) によって得られた個々の特徴表現ごとに、その出力に必要なパラメータ (例えば、局面の優劣に関する概念であれば「激指」によって計算されたその局面の評価値) を決定する。

実際の解説文 $\hat{s}$ の生成は、出力すべき特徴表現 $(\boldsymbol{w})$ とパラメータ $(r_1, r_2, \ldots, r_k)$ を変数とし、解説文としての尤らしさを表す言語モデルを目的関数とした最適化問題として定式化する。

$$\hat{s} = \underset{s}{\mathbf{argmax}}\, P_s(s|\boldsymbol{w}, r_1 r_2 \cdots r_k)$$

この式の右辺の特徴表現 $(\boldsymbol{w})$ の推定と P_s の設計が本応募課題の最大のポイントである。

4. 研究成果

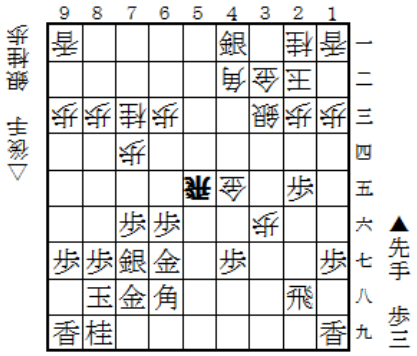
研究成果を各年度毎に分けて以下に記す。

- 26 年度: 将棋の解説文の生成には、将棋の解説文に対していくつかの言語処理がある程度の精度で行えることが前提となる。まず、必ず必要となる単語分割について正解データを作成して精度を測定した。結果、91%程度と低かったので、辞書の充実や学習データの追加を行い、97%程度まで高めることに成功した。これは、形態素解析器 KyTea の配布モデルとして公開している。また、23 種類の将棋用語を定義し、これらを自動認識するシステムを作成した。

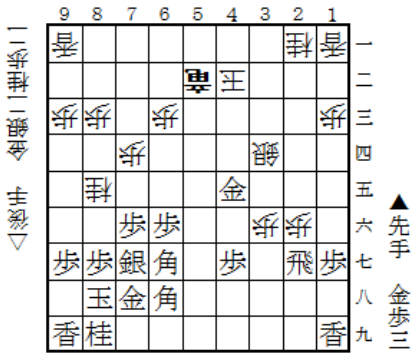
  次に、これらの言語処理を用いて単語に分割された解説文の単語と、局面のとの対応を自動的に取ることで、新たな局面に対してそれを表す特徴語を出力するモジュールを作成した。

  最後に、自動で単語に分割された大量の解説文から単語単位の確率的言語モデルを構築し、特徴語を入力として文を生成するモジュールを構築した。この結果、与えられた局面に対する文を生成することができるようになった。生成された文を精査した結果、精度は不十分であることがわかり、問題となる個所の改善について考案した。

- 27 年度: 将棋の局面とそれに対応するコメントを収集し、コーパスを作成した。 また、将棋に対して分野特有の固有表現を定義し、コーパス



解説「△７三の桂が動かないので先手が**指しやすい**ですね。」



解説「後手の玉が薄いので先手**優勢**だと思います。」

図 2: 解説の例

の一部、単語分割を行った 2,508 文に対して人手で固有表現をアノテーションした。 さらに、そのデータと BCCWJ を使って学習した KyTea と PWNER によって固有表現タグをコーパスの残り、742,286 文に付与した。

これらのコーパスを用いてニューラルネットワークによる局面と特徴語の自動対応、および言語モデルによる自動解説システムを構築した。局面に対する感想を述べられるようになった。

また、盤面を参照することで固有表現の認識精度を向上させる方法を提案し、実験的に効果を確認した。

- 28 年度: 与えられた盤面から先読みを行った結果得られる盤面に対して、すでに可能となっていた盤面状態からの解説文生成を応用し、解説を生成することを行った。このとき、最善手や次善手に加えて、解説されやすい手順も解説対象とすることを提案した。具体的には、解説されやすさをモデル化し、実際の解説からモデルのパラメータを学習して、手順を選択した。

  自動解説に加えて、言語のキーワードによる局面検索の手法を提案し、情報検索のトップ会議である ACM SIGIR 2017 に採択された。この研究では、解説文生成において構築した、シンボルグラウンディングモジュールを用いている。

  また、本研究テーマを通して作成した将棋の固有表現コーパスを用いて、局面を参照する固有表現認識器を提案し、局面を参照しない従来の固有表現認識に対する優位性を示した。実世界を参照する自然言語処理の先駆的研究として評価され、言語処理のトップ会議である ACL 2016 に採択され、発表を行った。

5. 主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

文献は以下の URL にて公開されている。

`http://www.ar.media.kyoto-u.ac.jp/mori/research/public/main.html`

〔雑誌論文〕（計　3 件）

1. シンボルグラウンディングによる分野特有の単語分割の精度向上
   友利 涼, 亀甲 博貴, 二宮 崇, 森 信介, 鶴岡 慶雅
   自然言語処理, Vol.24, No. 3, 2017. (to appear)
2. A Comparative Study of Dictionaries and Corpora as Methods for Language Resource Addition
   Shinsuke Mori, Neubig Graham
   LRE, Vol.50, pp.245-261, 2016.
3. 対数線形言語モデルを用いた将棋解説文の自動生成
   亀甲 博貴, 三輪 誠, 鶴岡 慶雅, 森 信介, 近山 隆
   情報処理学会論文誌, Vol.55, No.11, pp.2431-2440, 2014.

〔学会発表〕（計　16 件）

1. Game State Retrieval with Keyword Queries
   Atsushi Ushiku, Shinsuke Mori, Hirotaka Kameko, Yoshimasa Tsuruoka
   ACM SIGIR, 2017 年 8 月 7 日-11 日, 新宿, 2017. (to appear)
2. 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
   言語処理学会第 23 回年次大会, 2017 年 3 月 11 日-15 日, 筑波, 2017.
3. 実現確率に基づく解説すべき指し手の推定
   亀甲 博貴, 森 信介, 鶴岡 慶雅
   第 21 回ゲームプログラミングワークショップ, 2016 年 11 月 4 日-6 日, 箱根, pp.28-35, 2016.
4. Domain Specific Named Entity Recognition Referring to the Real World by Deep Neural Networks
   Suzushi Tomori, Takashi Ninomiya, Shinsuke Mori
   ACL, 2016 年 8 月 7 日-12 日, Berlin (Germany), 2016.
5. A Japanese Chess Commentary Corpus
   Shinsuke Mori, John Richardson, Atsushi Ushiku, Tetsuro Sasada, Hirotaka Kameko, Yoshimasa Tsuruoka
   LREC, 2016 年 5 月 22 日-25 日, Portoroz (Slovenia), pp.1415-1420, 2016.
6. 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
   言語処理学会第 22 回年次大会, 2016 年 3 月 7 日-11 日, 仙台, 2016.
7. 特徴語との自動対応によるゲーム局面の検索
   牛久 敦, 森 信介, 亀甲 博貴, 鶴岡 慶雅
   第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2016 年 2 月 29 日-3 月 2 日, 福岡, 2016.

8. Can Symbol Grounding Improve Low-Level NLP? Word Segmentation as a Case Study
   Hirotaka Kameko, Shinsuke Mori, Yoshimasa Tsuruoka
   EMNLP, 2015 年 9 月 17 日-21 日, Lisbon (Portugal), 2015.

9. Learning a Game Commentary Generator with Grounded Move Expressions
   Hirotaka Kameko, Shinsuke Mori, Yoshimasa Tsuruoka
   IEEE CIG, 2015 年 8 月 31 日-9 月 2 日, Taichun (Taiwan), 2015.

10. Named Entity Recognizer Trainable from Partially Annotated Data
    Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, Yoko Yamakata
    PACLING, 2015 年 5 月 19 日-21 日, Bali (Indonesia), 2015.

11. 将棋解説文のグラウンディングのための指し手表現と局面状態の対応付け
    亀甲 博貴, 森 信介, 鶴岡 慶雅
    第 19 回ゲームプログラミングワークショップ, 2014 年 11 月 7 日-9 日, 箱根, pp.202-209, 2014.

12. Language Resource Addition: Dictionary or Corpus?
    Shinsuke Mori, Graham Neubig
    LREC, 2014 年 5 月 26 日-31 日, Reykjavik (Iceland), pp.1631-1636, 2014.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）
○取得状況（計　0 件）

〔その他〕

- ゲーム解説コーパス
  `http://www.ar.media.kyoto-u.ac.jp/data/game/`
- 形態素解析器
  `http://www.phontron.com/kytea/index-ja.html`

6. 研究組織

(1) 研究代表者
森 信介 (Shinsuke Mori)
京都大学・学術情報メディアセンター・教授
研究者番号：90456773

(2) 研究分担者
鶴岡 慶雅 (Yoshimasa Tsuruoka)
東京大学・大学院工学系研究科・准教授
研究者番号：50566362