# 打者成績からのイニング速報の自動生成

村上 聡一朗 † 笹野 遼平 ‡ 高村 大也 ‡ 奥村 学 ‡
† 東京工業大学 大学院総合理工学研究科 ‡ 東京工業大学 精密工学研究所
† murakami@lr.pi.titech.ac.jp ‡{sasano,takamura,oku}@pi.titech.ac.jp

## 1 はじめに

野球の試合経過を速報するウェブサイトは多く存在する. これらのサイトでは, 一般的に, 図 1 のように各イニングの打者成績, スコアボードといったデータと人手で書かれた速報が記載されている. このうち速報は, 試合経過を知る記者が人手で記述したものであり, 試合経過に応じてリアルタイムに更新する作業はコストが大きいと考えられる. そこで本研究では, 打者成績からイニング速報を自動生成することを考える.

統計データや履歴情報等から, その状況を説明する文を生成する研究はいくつか行われている. たとえば, Belz[2] は, 天気予報に関する情報から天気予報のテキストの自動生成に取り組んでおり, 亀甲ら [5] は, コンピュータ将棋プログラムの局面解析情報を用いて将棋の解説文の自動生成を行っている. また, 野球の試合の要約記事の生成に関する研究として, Allen らの研究 [1] や Oh らの研究 [4] が挙げられる. Allen らは, 野球のスタッツ [1] と打撃内容の要約から, 新聞記者が書いたような試合の要約記事の自動生成を行っている. Oh らは, 野球のスタッツと打撃内容の要約から, 各チームの視点から見た試合の要約記事の自動生成を行っている. これらの研究では, 試合のスタッツや人手で書かれた打席毎の要約文などから試合全体の要約記事を生成することを目的としているのに対し, 本研究では, 打者成績からイニング速報を生成することを目的とする.

## 2 提案手法

提案手法の概要を図 2 に示す. 本研究では, 速報中のそれ以上細かく分割できない一連の事象のことを**イベント**と呼び, イニング速報生成を, 打者成績系列からイベントの系列を予測する系列ラベリング問題として扱う [2]. 提案手法は大きく予測モデルの学習フェーズと, 速報生成フェーズの 2 つに分けられる.

学習フェーズでは, まず (1) 人手で書かれたイニング速報を複数のイベントに分割した後, 打者名を手掛かりとし各イベントと打者成績の対応付けを行う. 続いて (2) 各イベントに含まれる打者名等を汎化することで各イベントをテンプレート化し, 最後に (3) テンプレート化されたイベントテンプレートの系列の予測

[1]打率, 打点数などの打撃成績, 投手の成績.
[2]本研究では, 打者成績からイニング速報の生成を行うことを目的とする. このため, 打者成績と対応しない走者, 投手, 記録などを表すイベントは生成の対象とはしない.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | R |
|---|---|---|---|---|---|---|---|---|---|---|
| ソフトバンク | 1 | 0 | 2 | 0 | 3 | 0 | 2 | 0 | 1 | 9 |
| 楽天 | 0 | 0 | 4 | 0 | 0 | 0 | 2 | 0 | 0 | 6 |

ソフトバンク・打者成績 (5回表)

| | | | | | | |
|---|---|---|---|---|---|---|
| 細川 | 明石 | 本多 | 柳田 | 内川 | 李大浩 | 松田 |
| 左中3 | 中安 | 中2 | 右安 | 右飛 | 二飛 | 三ゴ |

5回表

この回から2人目・松井が登板。先頭・細川が三塁打で出塁すると、明石のタイムリーで1点。さらに本多の二塁打で二・三塁とすると、柳田のタイムリーで2点。この回計3点を追加。

図 1: 打者成績とイニング速報の例

モデルの学習を行う. 速報生成フェーズでは, 入力された打者成績系列に, 学習フェーズで生成された予測モデルを適用することで, イベントテンプレートの系列を生成し, 打者名等を補うことで最終的なイニング速報を生成する. 以降では, 学習フェーズ中の (1) イベントと打者成績の対応付け, (2) イベントのテンプレート化, (3) 打者成績とイベントの対応関係の学習についてそれぞれ詳細を説明する.

### 2.1 イベントと打者成績の対応付け

野球のイニング速報では, 図 2 中の「岡、中島卓が連続安打で一・三塁とすると、西川の内野ゴロ間に 1 点を返す。」のように, 1 つの文に 2 つ以上のイベントが含まれる場合がある. そこで, 文中の読点, 動詞の連用形, 接続助詞, 並立助詞を分割箇所とし, 速報中の各文をイベントごとに分割する. ただし, 分割箇所の直前に選手名がある場合は例外とし分割しない. これは, 「先頭明石、柳田の連続安打、内川の犠打で 1 死二・三塁とする。」という文があった場合, 読点により分割されて「先頭明石」というイベント文が生成されることを防ぐためである.

続いて, 各イベントに含まれる打者名を手掛かりとして, イベントと打者成績を対応付ける. 具体的には, 図 2 のように, 打者成績「安打, 安打」を打者名「岡, 中島卓」を手掛かりとして, イベント文「岡、中島卓の連続安打で一・三塁とする」に対応付けを行う.

さらに, 「三者凡退」,「攻撃終了」,「試合終了」などのイニングの終わりに記述されるイベントを生成できるようにするため, イニングの最後の打者成績の後に END ノードを追加し, これらのイベントに対応付ける. 表 1 に使用した END ノードの種類を示す. END0 は, 三者凡退以外で, イニング中に打点が無い時に追加され, END1 はイニング中に打点が有る時に追加される. また, 「後続倒れ無得点で試合終了。」や「三者

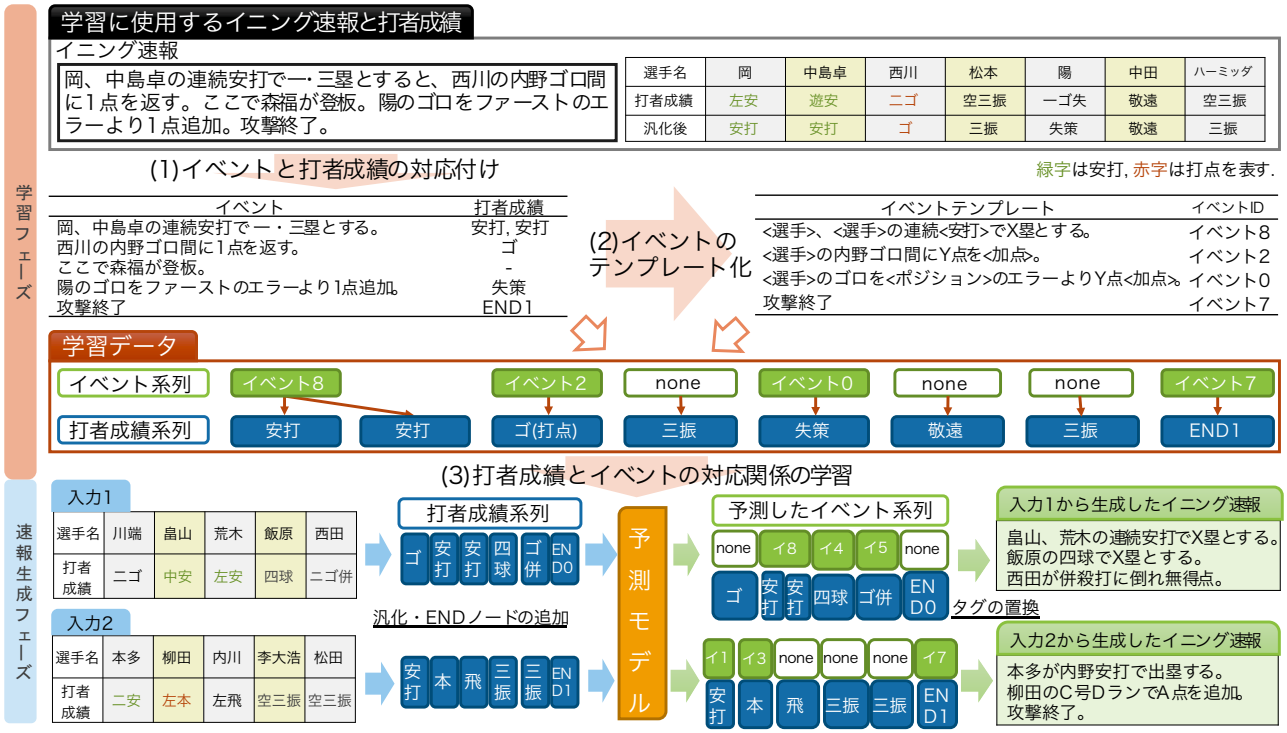


図 2: 提案手法の概要

表 1: END ノードの種類

| END0 | END1 | END2 | END3 |
|---|---|---|---|
| 打点無し | 打点有り | 試合終了 | 三者凡退 |

凡退。」といったイベントを予測するために, 打点の有無に関わらず試合終了のイニングに関しては END2 を, 三者凡退で終わったイニングに関しては END3 を追加する. たとえば, 図 2 の学習フェーズでは, 打点が入ったイニングであることから, 打者成績系列の末尾に END1 が追加されている.

## 2.2 イベントのテンプレート化

イベントごとに分割された過去のイニング速報のイベント文を自動生成の際に利用できるようにするため, イベント文のテンプレート化を行う. 具体的には, イベント文中の選手名や打者成績の情報, 塁, アウト数などを同定し, < 選手名 > や < アウト数 > などの汎化タグで置換することで汎化を行う [3]. たとえば, 図 2 では, 獲得したイベント文「岡、中島卓が連続安打で一・三塁とする。」から打者名や塁, 打者成績の情報のある箇所を同定し, 「< 選手 >、< 選手 > が連続 < 安打 > で X 塁とする。」というイベントテンプレートが生成されている.

また, イベント文と同様に, 打者成績もより一般的な打者成績に変換する. 具体的には,「左安」,「二ゴ」,「一ゴ失」などの打者成績から, 打球の行方を表す情報等を削除し, それぞれ「安打」,「ゴ」,「失策」に変換する. これは, イニング速報では「1 死から明石が安打で出塁する」などのように, 打球の行方に関する情報が含まれていない場合があること, および, 失策など比較的出現頻度の少ない打者成績をそのまま使用すると, 打者成績とイベントの対応関係がスパースになると考えられることが理由である.

続いて, 各テンプレートにイベント ID を付与する. まず, テンプレートのクラスタリングを行わない場合は, 完全に同一のテンプレートのみを 1 つにまとめた上で, 各テンプレートに固有のイベント ID を付与する. しかし, いくつかのテンプレートはほぼ同じ内容を表していると考えられることから, テンプレートをクラスタリングすることを考える. 具体的には, テンプレート間の類似度を, 各テンプレートに含まれる内容語および汎化タグの tf・idf 値で構成されるベクトルの余弦類似度で定義し, k-means 法によるクラスタリングを行い, 生成されたクラスタごとに固有のイベント ID を付与する. クラスタリングを行う場合, イニング速報の生成時には, クラスタの中心に近い 5 つのテンプレートの中から, 学習データ内で最も出現頻度が高いものをクラスタを代表するイベントテンプレートとして決定し, 代表イベントテンプレートを用いてイニング速報の生成を行う. 実験では, クラスタリングの効果を検証するために, クラスタリングを行った場合, 行わない場合のそれぞれに対して評価を行う.

## 2.3 打者成績とイベントの対応関係の学習

本研究では, 未知の打者成績系列からイニング中に起きたイベントを説明するイベントテンプレート系列を予測し, 各イベントテンプレートを用いてイニング速報を生成する. これは, 打者成績系列が与えられた

[3] 他にも点数やホームランに関する情報 (X 号, Y ラン) などの数字の汎化も行う.

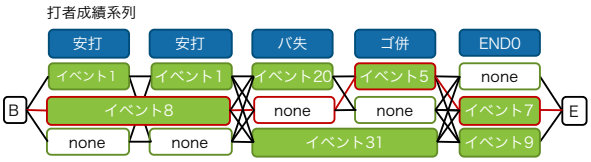


図 3: 打者成績系列に対するイベントラティス

時に, イベントテンプレート系列を予測する系列ラベリング問題として考えることができる. そこで本研究では, 系列ラベリング問題のための識別モデルである Conditional Random Fields (CRF) を用いて打者成績系列に対するイベントテンプレート系列の学習を行う.

イベントの中には, 図 2 中の「< 選手 >、< 選手 > が連続安打で X 塁とする。」のように, 複数の打者成績と対応付くイベントが存在することから BIO タグを用いた Linear-Chain CRF と semi-Markov CRF を用いる 2 つの手法を提案する.

#### 2.3.1 Linear-Chain CRF を用いた学習

Linear-Chain CRF を用いる手法では, 固有表現抽出などに使われる BIO タグを使用する. BIO タグのそれぞれの意味を以下に示す.

**B** その打者成績が対応するイベントの開始位置である
**I** その打者成績が対応するイベントの中にある
**O** その打者成績に対応するイベントが無い

2.2 節で生成したイベントテンプレートの異なり数が $N$ であった場合, BIO タグのタイプ数は $2N+1$ となる.

#### 2.3.2 semi-Markov CRF を用いた学習

semi-Markov CRF は, 入力系列に対する最適な境界と各境界部分に対するラベルを学習するモデルである. そのため, 単語境界が明示的でない日本語の形態素解析などに応用されている [3].

図 3 に入力の打者成績系列に対するイベントラティスの例を示す. この例では, 「安打, 安打, バ失, ゴ併, END0」という打者成績の入力に対し, 「イベント 8, none, イベント 5, イベント 7」というイベントテンプレート系列が出力される.

## 3 実験

### 3.1 実験設定

実験には, エキサイトベースボール [4] から収集した 2013 年から 2015 年のプロ野球の試合 20,300 イニング分のイニング速報と打者成績を利用した. この内, 20,000 イニングを学習データ, 200 イニングを評価データ, 100 イニングを開発データとして使用した. また, クラスタリングには Scikit-learn[5] の kmeans++を使用し, クラスタ数 $N$ は 200 とした. 形態素解析とイベント-打者成績間の対応関係の学習には, MeCab[6] を用いた.

表 2: 設定したルール

| 番号 | 条件 | 出力イベント |
|---|---|---|
| 1 | 打点有りの安打か | タイムリー |
| 2 | 凡打のみで打点が 3 人 | 三者凡退 |
| 3 | 連続する安打があるか | 連続安打 |
| 4 | 連続する四球があるか | 連続四球 |
| 5 | 連続する死球があるか | 連続死球 |
| 6 | 打者成績が安打 | 安打 |
| 7 | 打者成績が四球 | 四球 |
| 8 | 打者成績が死球 | 死球 |
| 9 | 打者成績が併殺打 | 併殺打 |
| 10 | 打者成績が犠打 | 犠打 |
| 11 | 打点有りの犠打か | タイムリー犠打 |
| 12 | 打者成績がバント失敗 | バント失敗 |
| 13 | 最後のイニングか | 試合終了 |
| 14 | 安打の後続が凡打で終了 | 後続倒れる |
| 15 | 打点が入ったイニングの終了時 | 攻撃終了 |
| 16 | 出塁イベント後の本塁打 | X ランホームラン |
| 17 | 出塁イベントが無い時の本塁打 | ソロホームラン |
| 18 | 打点有りのゴロか | ゴロ間に打点 |
| 19 | 守備のエラーで出塁したか | 失策 |

本研究では, 打者成績だけを用いてイニング速報を生成するため, 打者成績に関連付かないイベントは学習データから除外している. また,「出現回数が少ないテンプレートが使用されることは少ない」という仮定に基づき, データセット内で出現回数が 3 以下のテンプレートを含む事例は学習データとして使用しない.

提案手法の有用性を評価するために, 以下の 2 つをベースライン手法として使用した.

**ルールベース (rulebase)** 100 件の開発データセットを参考に, 打者成績系列からイニング速報を生成するための 19 のルールを設定した. 表 2 に設定したルールを示す.

**1 対 1 学習 (one2one)** 提案手法では, 連続する打者成績が 1 つのイベントに対応付くことに着目して, BIO タグを用いた Linear-Chain CRF や semi-Markov CRF を用いてイベント-打者成績間の対応関係を学習した. これに対し 1 対 1 学習手法では, 多対 1 の対応関係を考慮せず, 1 対 1 の対応関係だけを学習する. 具体的には, 多対 1 の対応関係がある事例を学習データから除き, Linear-Chain CRF で学習を行った.

実験では, クラスタリングを行わない場合の各提案手法 (BIO w/oC, semi w/oC), テンプレートのクラスタリングを行う場合の各提案手法 (BIO w/C, semi w/C), 2 つのベースライン手法 (rulebase, one2one), 人手で書かれたイニング速報 (HUMAN) の計 7 つのイニング速報と打者成績のペアを, 手法名を伏せて 5 人の評価者に提示し [7], 4 段階の評価を行った. 表 3 に提案手法と比較手法の概要, 表 4 に 4 段階の評価指標を示す. また, 本研究では, 打者成績だけを用いてイニング速報を生成するため, 打者成績から予測できないタグ (走者数, 打点数等) を埋めることができない. そのため, 評価者には, タグに正しい情報が入っているもの

[4] http://www.tbs.co.jp/baseball/top/main.html
[5] http://scikit-learn.org/
[6] http://taku910.github.io/mecab/
[7] 同一の速報が生成された場合は 1 つにマージし提示した.

表 3: 提案手法と比較手法の概要

| 手法名 | | 概要 |
|---|---|---|
| 提案手法 | BIO w/C | クラスタリングを行い, Linear-Chain CRF を用いて学習 |
| | semi w/C | クラスタリングを行い, semi-Markov CRF を用いて学習 |
| | BIO w/oC | クラスタリングを行わず, Linear-Chian CRF を用いて学習 |
| | semi w/oC | クラスタリングを行わず, semi-Markov CRF を用いて学習 |
| ベースライン手法 | rulebase | ルールより生成したイニング速報 |
| | one2one | 対応関係が 1 対 1 のみの事例を Linear-Chain CRF で学習 |
| HUMAN | | 人手で書かれたイニング速報 |

表 4: 評価指標の概要

| 点数 | 評価指標 |
|---|---|
| 4 | 打者成績を適切に説明できており, 理想的なイニング速報である. |
| 3 | 内容に誤りは含まれないが, 含むべき情報が欠けていたり, 冗長な印象を受ける. |
| 2 | 細かい誤りや不必要な情報を含んでいたり, 重要な情報が欠けている. |
| 1 | 重大な誤りを含んでいたり, 重要な情報がまったく含まれていないなど, イニング速報として不適切である. |

として評価するように説明を行った. 人手で書かれたイニング速報に対しては, 提案手法やベースライン手法と条件を等しくするためにイベント分割を行い, 打者成績から予測できない箇所の汎化を行ったものを評価対象としている. また, 有意差検定には, 並べ替え検定を有意水準 0.05 で使用した.

### 3.2 実験結果

表 5 に手法ごとの各評価が選択された回数と平均評価値を示す. 実験では, semi w/C が最も高い平均評価値を得た. また, 提案手法と各ベースライン手法の平均評価値の差は有意であることが確認できたが, 人手によるイニング速報 (HUMAN) と提案手法の平均評価値の間には有意差は確認できなかった.

表 6 に各手法によって生成されたイニング速報の例と評価者によって付けられた評価値の平均を示す. 提案手法の例には, 提案手法の中で最も平均評価値の高かった semi w/C の結果を示している. 事例 A の one2one では, 打者成績とイベント間の多対 1 の対応関係を学習しないため, イニング速報中に連続安打という必要な情報が欠けており, 他の手法と比べて低い点数であることが分かる. また, 事例 B の rulebase では, 他の手法と比較すると内容に冗長な印象を受ける. また, 「押し出し四球」という重要な情報が欠けており, 他の手法と比べて低い点数となっている.

## 4 結論

本研究では, 過去の人手で書かれたイニング速報からテンプレートを生成し, 獲得したテンプレートと打者成績系列から予測したイベントテンプレート系列を基にイニング速報を自動生成する手法を提案した. 評価実験の結果, 提案手法は 2 つのベースライン手法よりも有意に高い評価を, 人手で書かれた正解と比べた場合も同等の評価を得ることができ, 提案手法の有用性が確認できた. 今後の課題としては, 本研究では扱わなかった投手成績等を考慮したイニング速報の生成, 自動生成したイニング速報を用いた試合全体の要約記事の生成などが考えられる.

表 5: 手法ごとの評価の内訳と平均点

| 手法名 | | 各評価の選択数 4 | 3 | 2 | 1 | 平均点 |
|---|---|---|---|---|---|---|
| 提案手法 | BIO w/C | 880 | 95 | 21 | 4 | 3.851 |
| | semi w/C | 880 | 95 | 23 | 2 | **3.853** |
| | BIO w/oC | 852 | 113 | 27 | 8 | 3.809 |
| | semi w/oC | 868 | 108 | 22 | 2 | 3.842 |
| ベースライン手法 | rulebase | 735 | 222 | 33 | 10 | 3.682 |
| | one2one | 679 | 202 | 103 | 16 | 3.544 |
| HUMAN | | 866 | 102 | 22 | 10 | 3.824 |

表 6: 提案手法, 比較手法のイニング速報の例

| 事例A | 山田 | 上田 | 川端 | 雄平 | 畠山 | デニング | 森岡 |
|---|---|---|---|---|---|---|---|
| | 空三振 | 投ゴ | 左安 | 中安 | 中安 | 左安2 | 投ゴ |

| 手法名 | 平均点 | イニング速報 (事例 A) |
|---|---|---|
| semi w/C | **3.80** | 2 死から川端、雄平の連続安打で B 塁とする。畠山のタイムリーで A 点を追加。デニングのタイムリーツーベースで A 点を追加。 |
| one2one | 3.00 | 2 死から川端の安打。雄平の安打で B 塁とする。畠山のタイムリーで A 点。デニングのタイムリーツーベースで A 点。 |
| rulebase | **3.80** | 川端と雄平の連続安打などで 2 死 B 塁とする。畠山のタイムリーで A 点を追加。デニングのタイムリーツーベースで A 点を追加。攻撃終了。 |
| HUMAN | 3.60 | 2 死から川端、雄平の連続安打で B 塁とする。畠山のタイムリーで A 点。デニングのタイムリーツーベースで A 点。 |

| 事例B | 秋山 | 渡辺 | 浅村 | 中村 | メヒア | 栗山 | 熊代 | 炭谷 | 永江 | 秋山 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 左安 | 死球 | 四球 | 四球 | 右安 | 一ゴ | 四球 | 一邪飛 | 死球 | 中飛 |

| 手法名 | 平均点 | イニング速報 (事例 B) |
|---|---|---|
| semi w/C | 3.00 | 秋山の安打。渡辺の死球。浅村の四球などで無死 B 塁とする。中村の押し出し四球で A 点。メヒアのタイムリーで A 点。栗山の内野ゴロ間に A 点。2 死から永江の死球。 |
| one2one | 3.00 | 秋山の安打。渡辺の死球。浅村の四球などで無死 B 塁とする。中村の押し出し四球で A 点。メヒアのタイムリーで A 点。栗山の内野ゴロ間に A 点。永江の死球などで 2 死 B 塁とする。 |
| rulebase | 2.40 | 無死から秋山が安打で出塁する。無死から渡辺が死球で出塁する。無死から浅村が四球で出塁する。メヒアのタイムリーで A 点を追加。栗山のファーストゴロ間に A 点を追加。1 死から熊代が四球で出塁する。2 死から永江が死球で出塁する。攻撃終了。 |
| HUMAN | **3.40** | 秋山の安打。渡辺の死球。浅村の四球で B 塁とする。中村の押し出し四球で A 点。メヒアのタイムリーで A 点。栗山の内野ゴロ間に A 点。熊代の四球。永江の死球などで 2 死 B 塁とする。攻撃終了。 |

**謝辞** 本研究は JSPS 科研費 26280080 の助成を受けた.

## 参考文献

[1] Nicholas D. Allen, John R. Templon, Patrick Summerhays McNally, Larry Birnbaum, and Kristian Hammond. Statsmonkey: A data-driven sports narrative writer. In *AAAI Fall Symposium*, 2010.

[2] Anja Belz. Probabilistic generation of weather forecast texts. In *Proc. of NAACL-HLT'07*, pp. 164–171, 2007.

[3] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP'04*, Vol. 4, pp. 230–237, 2004.

[4] Alice Oh and Howard Shrobe. Generating baseball summaries from multiple perspectives by reordering content. In *Proc. of INLG'08*, pp. 173–176, 2008.

[5] 亀甲博貴, 三輪誠, 鶴岡慶雅, 森信介, 近山隆. 対数線形言語モデルを用いた将棋解説文の自動生成. 情報処理学会論文誌, Vol. 55, No. 11, pp. 2431–2440, 2014.