テンプレートの自動生成によるイニングの要約文生成

田川裕輝† 嶋田和孝‡
† 九州工業大学大学院 情報工学府 ‡ 九州工業大学大学院 情報工学研究院
{y_tagawa, shimada}@pluto.ai.kyutech.ac.jp

1 はじめに

スポーツの分野で特に人気の高い野球は試合の速報がWeb などで配信されている．速報には打席ごとに更新されるテキスト速報とイニング終了時にまとめて更新されるイニング速報がある．テキスト速報とイニング速報の例を表 1 に示す．テキスト速報は各打席の打撃内容や選手交代，ポジションの変更などの情報を羅列しており，読みやすいとは言い難い．一方，イニング速報はこれらの情報を文の形式でまとめた速報であるが，イニング中の各打席を網羅的に説明した文であり，比較的長い文である．そのため，短いイニング速報を生成することで，読み手の負担を減らすことができる．本研究では，テキスト速報からイニングの情報を簡潔に説明する文の生成に取り組む．

本研究ではテンプレート型生成手法を用いて文を生成する．テンプレート型生成手法とは，生成する文の雛形となるテンプレートを事前に用意し，テンプレートに必要な情報を補完することで文を生成する手法である．この手法では，文法的に正確な文が生成できるといった利点があるが，テンプレートを用意するコストが大きいといった欠点がある．そこで本研究では，テンプレートを自動で生成する文生成手法を提案する．

提案手法により，実際に表 1 のテキスト速報から生成された文の例を以下に示す．

1. 陽のタイムリーヒットで 1 点。杉谷のタイムリースリーベースで 3 点。この回計 4 点で攻撃終了。

この文は陽と杉谷の 2 打席に言及した文であることがわかる．このように，本研究ではイニング中の全打席ではなく，注目すべき打席に焦点をあてたシンプルな文を生成する．

一方で，1 の文だけでは，試合の流れなどが理解しづらい．そこで,“走者一掃の”や“打者一巡の猛攻で”といった試合の流れを考慮したフレーズ (Game-changing Phrase, 以下 GP) を含む，より洗練された文の生成も目指す．

2. 同点で迎えた 5 回、陽のタイムリーヒットで 1 点。杉谷の走者一掃のタイムリースリーベースなど打者一巡の猛攻で 4 点を挙げ勝ち越しに成功。

この文は 1 の文を洗練化したものである．

まとめると，本研究の貢献は

- テンプレート自動生成による文生成手法の提案
- シンプルな文と洗練された文の両方を生成

の 2 つである．

表 1: 2016 年 9 月 7 日 日本ハム対ロッテ戦 9 回表 のテキスト速報とイニング速報の例

テキスト速報
中田　　外角の真っ直ぐを打つもレフトフライ 1 アウト 田中賢　一度もバットを振ることなくフォアボールを選ぶ 1 塁 投手交代:古谷→大嶺祐 一塁走者田中賢介:盗塁成功 2 塁 レアード　鈴木 (遊) のファンブルにより出塁する 1,3 塁 陽　　1 アウト 1,3 塁の 2-2 から勝ち越しのタイムリーヒット! ロ 2-3 日 1,2 塁 大野　1 アウト 1,2 塁から内角のストレートを打つもセンターフライ 2 アウト 中島卓也　ランナー 1,2 塁からレフトへのヒットを放つ 満塁 岡→代打:杉谷 杉谷　2 アウト満塁からセンターへのタイムリースリーベース！ ロ 2-6 日 3 塁 西川　粘りを見せて 8 球目にフォアボールを選ぶ 1,3 塁 一塁走者西川:盗塁成功 2,3 塁 大谷　カウント 1-2 から外角のフォークに空振り三振でバッターアウト 3 アウトチェンジ
イニング速報
1 死から田中賢の四球で出塁すると、ここで 2 人目大嶺祐が登板。田中賢の盗塁、レアードがショート鈴木のエラーで出塁し一三塁とすると、陽のタイムリーで 1 点。中島卓の安打で満塁とすると、代打杉谷の走者一掃のタイムリースリーベースで 3 点。この回計 4 点をとり、勝ち越しに成功。

2 関連研究

スポーツニュースを自動で生成する研究 [1–7] は試合のスタッツや選手の成績などのデータを入力とした研究 [1,2] や SNS，テキストコメンタリーに投稿されるテキストを入力とした研究 [3–5]，テキスト速報などのニュース記事を入力とした研究 [6,7] など盛んに行われている．

村上ら [1] は，各イニングの打者成績の系列を入力とし，表 2 のようなイニング速報を自動で生成している．村上らの目的はイニングの情報を網羅した速報を生成することが目的であり，本研究と目的が異なる．

岩永ら [6] は，野球のテキスト速報を入力とし，試合ダイジェストを自動で生成している．また，Oh ら [7] は，野球に関するニュース記事からホームチーム視点の記事とアウェイチーム視点の記事を自動で生成するアルゴリズムを提案している．岩永ら，Oh らの研究は試合ダイジェストの生成を目的としており，本研究と目的が異なる．

Robin ら [2] は，バスケットボールのボックススコアを入力として，試合ダイジェストを生成している．また，得点シーンを説明する文では“score”というシンプルな表現だけでなく“pump in”や“fire in”，“dish out”など多様な表現を使用することで，より自然で柔軟な文を生成している．本研究でも，このように多様な表現を使用し，より洗練された文の生成を目指す．

表 2: 2016 年 8 月 3 日 日本ハム対ロッテ戦 5 回表 のテキスト速報から抽出したデータ

打者名	アウト数	打撃前出塁状況	打撃内容	打撃後出塁状況	得点数	打席結果
中田	0 アウト	0 塁	レフトフライ	0 塁	0	得点なし
田中賢	1 アウト	0 塁	フォアボール	1 塁	0	得点なし
田中賢	1 アウト	1 塁	盗塁	2 塁	0	得点なし
鈴木 (遊)	1 アウト	2 塁	ファンブル	1,3 塁	0	得点なし
陽	**1 アウト**	**2,3 塁**	**タイムリーヒット**	**1,2 塁**	**1**	**勝ち越し**
大野	1 アウト	1,2 塁	センターフライ	1,2 塁	0	得点なし
中島卓也	1 アウト	1,2 塁	ヒット	満塁	0	得点なし
杉谷	**2 アウト**	**満塁**	**タイムリースリーベース**	**3 塁**	**3**	**追加点**
西川	2 アウト	3 塁	四球	1,3 塁	0	得点なし
西川	2 アウト	1,3 塁	盗塁	2,3 塁	0	得点なし
大谷	2 アウト	2,3 塁	空振り三振	2,3 塁	0	得点なし

表 3: 2016 年 8 月 7 日 西武対楽天 7 回裏の注目打席のデータとイニング速報

打者名	アウト数	打撃前出塁状況	打撃内容	打撃後出塁状況	得点数	打席結果
…	…	…	…	…	…	…
茂木	**1 アウト**	**1,3 塁**	**タイムリーヒット**	**1,2 塁**	**1**	**勝ち越し**
…	…	…	…	…	…	…
銀次	**1 アウト**	**満塁**	**タイムリーヒット**	**満塁**	**1**	**追加点**
…	…	…	…	…	…	…

イニング速報
先頭聖澤の二塁打、藤田の犠打で 1 死三塁とすると、嶋のスクイズで 1 点。その際にピッチャー高橋光の野選で二塁とすると、ここで 2 人目野田が登板。島内の安打で一三塁とすると、茂木のタイムリーで 1 点。ペゲーロの四球で満塁とすると、ここで 3 人目牧田が登板。ウィーラーの押し出し死球で 1 点。銀次のタイムリーで 1 点。この回計 4 点で攻撃終了。

3 提案手法

提案手法では前処理としてテキスト速報から表 2 のようなデータを抽出し，そのデータから文を生成する．まず，文として生成する注目打席を選択する．次に，テンプレートの自動生成手法を用いて，文を生成する．また，生成されたテンプレートに GP を融合することで，より洗練された文を生成する．

3.1 注目打席の選択

各打席ごとにスコアリングし，スコア上位 2 打席を注目打席とする．スコアリングには式 (1) を用いる．

$$\text{打席スコア} = \text{得点数} + \text{打撃内容スコア} \quad (1)$$

得点数とはその打席で得点した点数である．また，打撃内容スコアとはホームランやタイムリーヒットなどの重要な打撃内容に対しては高く，フライや三振などの打撃内容に対しては低くなるように定めたスコアである．例えば，表 2 のイニングでは，陽と杉谷の打席が注目打席として選択される．

3.2 テンプレートの自動生成と要約文生成

本研究では類似したデータを持つ打席同士は，その打席を説明する文の構造も類似していると考える．表 3 は表 2 とは別の試合のイニングで注目打席として選択された 2 打席のデータとそのイニング速報である．表 2 の注目打席 (陽と杉谷) のデータと表 3 の注目打席のデータは類似していることがわかる．つまり，表 3 のイニング速報中の類似打席 (茂木と銀次) に関するデータを注目打席のデータに置き換え，類似打席以外の打席に関する部分は文圧縮処理により削除することで，文法的に正確な注目打席を説明する文を生成することができると考える．このような考え方に基づき，注目打席と類似したデータを持つ類似イニングを検索し，テンプレートを自動生成したのち，要約文を生成する．

3.2.1 類似イニングの自動検索

検索する際には，注目打席とその他のイニングの注目打席の間でデータの一致率を計算し，最も一致率の高いものを類似イニングとする．

一致率は表 2 に示す打者名以外の 6 つの項目で計算する．ただし，打席結果の一致は絶対条件とし，打席結果の異なる打席の一致率は 0 とする．例えば，表 2 の陽の打席と表 3 の茂木の打席は打席結果が一致しており，それ以外の 5 項目のうち，打撃前出塁状況以外の 4 項目が一致している．よって一致率は 0.8 となる．同様に，表 2 の杉谷の打席と表 3 の銀次の打席は一致率が 0.4 となり，最終的に表 2 と表 3 の注目打席間の一致率は 1.2 となる．

3.2.2 文圧縮処理と文生成

次に，類似イニングのイニング速報中の類似打席に関する部分はスロット化，それ以外は文圧縮処理により削除することでテンプレートを生成する．

まず，類似イニングのイニング速報を動詞の連用形，句読点，並立助詞，接続助詞で分割する．ただし，分割点の直前が打撃内容語の場合は分割しない．

次に文圧縮処理を行う．表 4 に文圧縮処理から文生成までのフローを示す．分割された文のうち，選手名を含む文が文圧縮の対象となり，類似打席以外の打席に関するイベントを文圧縮処理により削除する．文圧縮では以下の 2 つの処理を順に行う．

文圧縮 1 類似打席以外の打者名からその打者の打撃内容までを削除する．またアウト数や投手の情報も削除する．

文圧縮 2 選手名を含んでいない文は文ごと削除する．

表 4: 文圧縮から文生成までのフロー

文圧縮 1	~~先頭聖澤の二塁打、藤田の犠打で1死~~三塁とすると、/ ~~嶋のスクイズで~~ 1 点。/ その際に~~ピッチャー高橋光の野選で~~二塁とすると、/ ここで 2 人目~~野田が登板。/ 島内の安打で~~一三塁とすると、/ 茂木のタイムリーで 1 点。/ ~~ペゲーロの四球~~で満塁とすると、/ ここで 3 人目~~牧田が登板。/ ウィーラーの押し出し死球~~で 1 点。/ 銀次のタイムリーで 1 点。/ この回計 4 点で攻撃終了。
文圧縮 2	~~で三塁とすると、/ で 1 点。/ その際にで三塁とすると、/ ここで 2 人目。/ で一三塁とすると、/~~ 茂木のタイムリーで 1 点。/ ~~で満塁とすると、/ ここで 3 人目。/ で 1 点。/~~ 銀次のタイムリーで 1 点。/ この回計 4 点で攻撃終了。
テンプレート化	[NAME1] の [ACTION1] で [SCORE1] 点。[NAME2] の [ACTION2] で [SCORE2] 点。この回計 [ALLSCORE] 点で攻撃終了。
要約文生成	陽のタイムリーヒットで 1 点。杉谷のタイムリースリーベースで 3 点。この回計 4 点で攻撃終了。

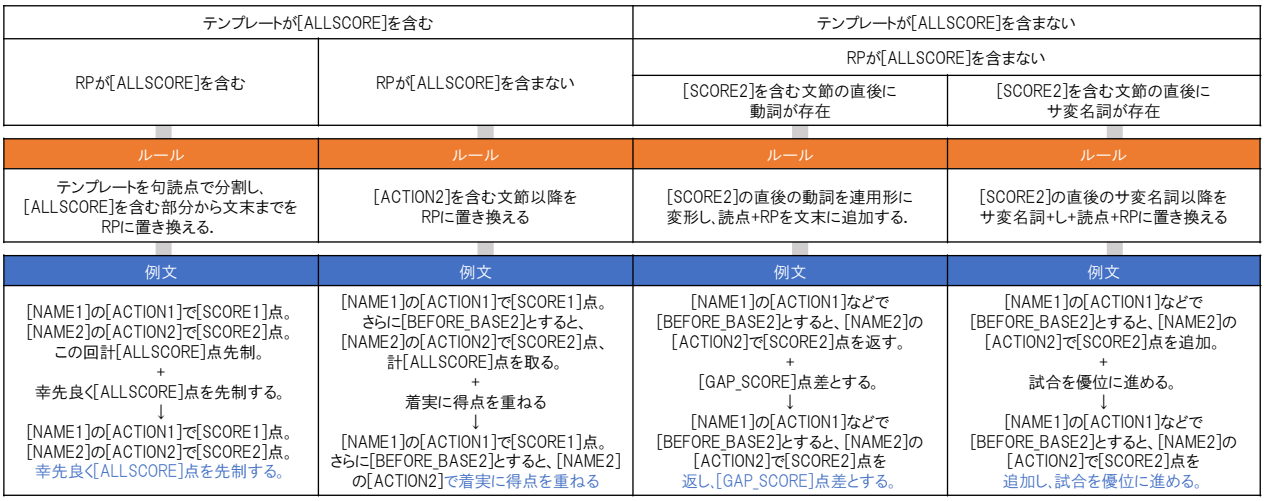

図 1: テンプレートと RP の融合フロー

例えば，文圧縮 1 の処理では表 4 に示すように，“ 先頭聖澤の二塁打 ” や “ 野田が登板 ” といったイベントを削除する．

文圧縮処理の後，打者名や打撃内容などをスロットとし，テンプレートを生成する．その後，注目打席のデータを補完することで，文を生成する．

3.3 Game-changing Phrase の融合

生成されたテンプレートに Game-changing Phrase(GP) を融合することでより洗練された文を生成する．

まず，GP は Yahoo!Sportsnavi[1] の試合ダイジェストから自動で獲得する．試合ダイジェストは勝利概要，得点シーン，投手概要，敗因の 4 文から構成されており，得点シーンに関する文から GP を自動で獲得する．得点シーンに関する文の例を以下に示す．

- 広島は 2 点を追う$_{\text{IP}}$9 回裏、ルナの 2 点$_{\text{AP}}$ 適時打が飛び出し、土壇場で試合を振り出しに戻す$_{\text{RP}}$。
- 中日は 両軍無得点のまま迎えた$_{\text{IP}}$7 回裏、吉見が値千金の$_{\text{AP}}$ 適時打を放ち、試合の均衡を破った$_{\text{RP}}$。

これらの文にはイニングを修飾する GP(緑文字部，Inning Phrase，以下 IP) や打撃内容を修飾する GP(赤文字部，Action Phrase，以下 AP)，同点や先制のようなイニング結果を表現する GP(青文字部，Result Phrase，以下 RP) が存在する．

これらの GP をパターンマッチにより獲得する．IP はチーム名を含む文節と回数の間に出現するものを獲得する．AP は打者名を含む文節と打撃内容の間に出現するものを獲得する．RP は “ 飛び出し ” や “ 放ち ” などのキーワードを含む文節から文末の間に出現するものを獲得する．

表 5: 獲得された GP とそのルールの例

GP	得点数	イニング数	結果
起死回生の	1 点以上	9 回	同点
[ALLSCORE] 点を挙げ、突き放す	2 点以上	4 回以前	追加点
試合を振り出しに戻す	1 点以上	6 回以降	同点
幸先良く先制する	1 点以上	1 回	先制
試合の均衡を破る	1 点以上	6 回以降	先制

次に，獲得された GP に対して，ルールを作成する．例えば，“ 土壇場で試合を振り出しに戻す ” という RP が使われている得点シーンは，全てホームランまたはタイムリーで同点に追いついた 9 回であることが確認できた．このように，各 GP に対して，イニング数やイニング結果，打撃内容などから簡単なルールを作成する．獲得された GP とそのルールの例を表 5 に示す．そして，注目打席のデータが作成したルールにマッチする場合，テンプレートに対して，該当する GP を融合する．

次に融合手法について説明する．IP と AP は，獲得した際に用いたルールを逆に適用し，融合する．要するに，IP はチーム名と回数の間に，AP は打者名と打撃内容の間に挿入する．RP は図 1 に示すフローで融合する．

最後にテンプレート中のスロットにデータを補完することで，最終的な文を生成する．

4 実験

2016 年 6 月 9 日から 2016 年 10 月 29 日までの 510 試合の Yahoo!Sportsnavi のテキスト速報と試合ダイジェスト，

[1] http://sports.yahoo.co.jp/

表 6: 提案手法により生成された文の例

GP なし	柳田のフォアボール、内川のホームランで 4 点。この回計 5 点を取り逆転に成功。
GP あり	1 点を追う$_{\text{IP}}$3 回、柳田の押し出しフォアボール、内川の 満塁$_{\text{AP}}$ ホームランなど 打者一巡の猛攻で一挙 5 点を奪い逆転に成功する$_{\text{RP}}$。
GP なし	清田のヒットなどで 1,3 塁とすると、角中のタイムリーツーベースで同点に追いつく。
GP あり	1 点ビハインドの$_{\text{IP}}$9 回、清田のヒットなどで 1,3 塁とすると、角中の 起死回生の$_{\text{AP}}$ タイムリーツーベースで 土壇場で試合を振り出しに戻す$_{\text{RP}}$。
GP なし	ボグセビックのフォアボールなどで 2 塁とすると、大城の三塁打で先制。
GP あり	両軍無得点のまま迎えた$_{\text{IP}}$6 回、ボグセビックのフォアボールなどで 2 塁とすると、大城の三塁打で 待望の先制点を挙げる$_{\text{RP}}$。

表 7: 評価指標

点数	評価指標
3	日本語として読みやすい文である.
2	助詞の省略や冗長な表現など細かい誤りはあるが, 比較的読みやすい文である.
1	日本語として読みにくい文である.

エキサイトベースボール [2] からイニング速報を収集した. 実際に, テキスト速報を入力として与え, 注目打席を選択し, そのデータから文を生成した. 表 6 に提案手法で生成された文の例を示す. 文中の緑, 赤, 青文字部分はそれぞれ融合された IP, AP, RP である.

4.1 評価

収集した 510 試合のデータうち, ランダムに 10 試合を選択し, 生成された 200 文に対して評価実験を行った.

テキスト速報を被験者に与えた上で, 生成された文を表 7 の評価指標から 3 段階で評価した. 評価結果の平均は 2.77 と高い評価であった. また, 生成された 200 文とイニング速報の文長を比較したところ, 提案手法の要約率は 0.57 と約 4 割短くできていることが確認できた.

4.2 考察

生成された文の評価結果は高い評価であったが, 実際には非文も生成されていた. 非文が生成された主な原因は文圧縮処理でのエラーであった. 例えば, イニング速報中には“XXX の四球, YYY の内野ゴロで走者が入れ替わり、”といった表現がある. この文から, XXX の打席に関するイベントをスロット化し, YYY の打席に関するイベントを削除する場合, 提案手法の文圧縮 1 の処理により, “YYY の内野ゴロ”が削除され, “[NAME] の [ACTION] で走者が入れ替わり、”といったテンプレートが生成される. 要するに, “走者が入れ替わる”という関係のない情報が残ったままテンプレートが生成される問題があった.

提案手法ではより洗練された文を生成するため, GP を試合ダイジェストから自動で獲得し, GP ごとにルールを作成した. 例えば, 表 5 に示した“起死回生の”という AP は 9 回に同点に追いついた場面で選択される. イニング数という情報は時間情報と捉えることができるため, サッカーであれば後半ロスタイムに同点に追いついた場面で選択することができる. このように, 野球以外のスポーツへの応用も可能であると考えられる.

一方で, GP は試合ダイジェスト中の得点シーンに関する文から自動で抽出し, ルールを作成したため, 得点シーン以外では, ルールにマッチする GP が存在しない. 例えば, 三者凡退のイニングで生成される文は“三者凡退”のみであり, 柔軟さに欠ける. そのため, “3, 4, 5 番のクリーンナップが並ぶも三者凡退”のように得点シーン以外で利用される GP の獲得が課題に挙げられる.

[2] http://www.tbs.co.jp/baseball/top/main.html

5 おわりに

本研究では, テンプレートの自動生成によるイニング情報を簡潔に説明する文の生成手法を提案した. また, 試合の状況を考慮したフレーズ (Game-changing Phrase) を文に組み込むことで, より洗練された文を生成した.

今後の課題としては新たな文圧縮処理手法の提案が挙げられる. また, 提案手法では注目打席を得点数と打撃内容からスコアリングにより選択したが, 直近数試合の打率や投手との相性などの個人成績をスコア式に組み込むことで, 注目選手に着目した文生成にも取り組みたい.

謝辞 この研究の一部は科研費 26730176 の助成を受けたものです.

参考文献

[1] 村上総一郎, 笹野遼平, 高村大也, 奥村学. 打者成績からのイニング速報生成. 言語処理学会第 22 回年次大会発表論文集, 2016.

[2] Jacques Robin. Revision-based generation of natural language summaries providing historical background. Ph.D. thesis, New York University, 1994.

[3] Yuuki Tagawa, Kauzutaka Shimada. Generating abstractive summaries of sports games from japanese tweets. In Proc.of ESKM ’16, pp. 82–87, 2016.

[4] J.Nichols, J.Mahmud, C.Drews. Summarizing sporting events using twitter. In Proc.of IUI ’12, pp. 189–198, 2012.

[5] Zhang Jianmin, Yao Jin-ge, Wan Xiaojun. Toward constructing sports news from live text commentary. In Proc.of ACL ’16, pp. 1361–1371, 2016.

[6] 岩永朋樹, 西川仁, 徳永健伸. テキスト速報を用いた野球ダイジェストの自動生成. 言語処理学会第 22 回年次大会発表論文集, 2016.

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