テキスト速報を用いた野球ダイジェストの自動生成

岩永 朋樹 † 西川 仁 ‡ 徳永 健伸 ‡
† 東京工業大学 工学部 ‡ 東京工業大学 大学院情報理工学研究科
iwanaga.t.ab@m.titech.ac.jp {hitoshi, take}@cs.titech.ac.jp

1 はじめに

スポーツの分野において，特に耳目を集める野球やサッカーなどでは，テレビ中継やニュースサイトなどを通じて試合の状況が随時配信されている．特に，各試合に対するテキストを媒体として行われる報道に着目すると，大きくわけて試合中に発生した何らかのイベントを随時簡潔に報道するものと，試合後にそれらの中で特に注目すべきイベントなどに焦点をあて試合の概要を報じるものの 2 種類が存在する．これらは現在ではいずれも人手で作成されているが，これらを仮に自動的に生成することができれば，人手をかけることによる費用および時間の削減を期待することができ，研究の余地がある．

本稿では特に野球の試合に関する報道を取り上げ，試合中に発生したイベントについて報じた簡潔なテキスト速報（以降，イベントと呼ぶ）を入力とし，それらを元にして当該試合の概要（以降，ダイジェストと呼ぶ）を生成する手法を提案する．

イベントの例を表 1 に示す．各イベントは試合中に生じた全てのイベントの通し番号，イニング，裏表の別，攻撃側のチーム名，打者，そして打席の内容からなる．打席の内容とは別に，盗塁や選手交代などもイベントに含まれる．

ダイジェストの例を表 2 に示す．ダイジェストは試合の簡潔な要約が記述されたテキストである．詳しくは後述するが，ダイジェストは大まかに分けて表 2 に示すように 4 つの構成要素からなることが多い．なお，実際のダイジェストはタグなどは付与されていない単なるテキストであり，表 2 に示す構成要素は著者らが本研究のために便宜的に付与したものである．

2 関連研究

スポーツにおける試合のダイジェストを生成する試みは自然言語生成 [5] の分野でなされてきた [1, 2]．これは，自然言語生成に際しては入力となる何らかの中間表現を陽に定義する必要があるが，スポーツの試合に関しては試合中に発生しうる出来事が規則で定義されており，これを中間表現として利用できること，加えてそのような中間表現と試合後に配信されるダイジェストの組が数多く用意されており，統計的な手法に基づくアプローチに親和的であることが理由であろう．

Barzilay らはアメリカン・フットボールの試合経過を入力として，当該試合の概要を生成する手法を提案している [1, 2]．Barzilay らの手法はそれぞれ，試合経過のうち特に重要なイベントの選択 [1] と，複数のイベントから単一の文の表層を生成する，自然言語生成における命題の凝集 [2] に焦点をあてている．本稿では重要なイベントの同定にはヒューリスティックを利用し，表層生成には規則を用いる．

生成の対象とするテキストの分野は異なるが，同様に中間表現からの生成を行うものとして Higashinaka ら [3] によるものがある．Higashinaka らは飲食店に関する評価文書から，評価に関する表現を抽出し，対話システムが評価に関する応答を生成する際にこれらの中から適切なものを選択することで応答を生成する手法を提案している．Higashinaka らの問題設定は内容選択が行われたあとの表層生成に焦点があてられており，本稿の問題設定とは異なっている．

3 提案手法

3.1 ダイジェストの構成

事前の調査によれば，表 2 に示す様に，ダイジェストは主に **(A) 勝利概要**，**(B) 打撃**，**(C) 投手**，**(D) 敗因**の 4 つの部分から構成されていることが多い．

(A) 勝利概要は勝利したチームに焦点をあて試合展開を簡潔に述べたものである．この (A) 勝利概要のパートは試合展開によってはダイジェスト中に存在しない場合もある．(B) 打撃は試合中の重要な打撃イベントについて述べたもので，例えば逆転ホームランなどが含まれる．(C) 投手は試合での継投や勝利投手の情報について述べたものである．(A)，(B)，(C) のパートは勝利チームを主語として記述されている．(D) 敗因は敗れたチームの敗因を述べたものであり，敗れたチームが主語として述べられる．

表 1: イベントの例

通番	イニング	裏表	チーム	打者	内容
2	1	表	楽天	牧田	ライトへのヒットを放つ 1 塁
10	2	表	楽天	-	一塁走者後藤:盗塁成功 2 塁
22	3	表	楽天	藤田	1 アウト 1,2 塁の 1-2 からセンターへの先制タイムリーツーベース！ロ 0-2 楽 2 塁
85	9	裏	ロッテ	-	ピッチャークルーズに代わって空振りを奪う松井裕がマウンドにあがる
…	…	…	…	…	…

表 2: ダイジェストの例

構成要素	内容
(A) 勝利概要	楽天が接戦を制した。
(B) 打撃	楽天は 3 回表、藤田の 2 点適時二塁打で先制する。
(C) 投手	投げては先発・戸村が 5 回 1 失点と試合をつくり今季 4 勝目。その後は青山、レイ、クルーズとつなぎ、最後は守護神・松井裕が締めた。
(D) 敗因	敗れたロッテは、打線が 11 残塁と拙攻が響き、連勝が 4 で止まった。

3.2 ダイジェストの生成

ダイジェストの生成は，(A) 勝利概要と (B) 打撃，(D) 敗因を別個の手法で生成し，それぞれの出力を結合することで行う．

本稿では (A) 勝利概要，(B) 打撃，(D) 敗因の 3 つの部分を生成の対象とし，(C) 投手は生成の対象外とする．本稿で実験のために利用するデータには打撃の情報しか含まれておらず，投手の情報は含まれていないため，このデータから (C) 投手について生成することが難しい．すなわち (C) 投手の部分を生成するための情報を得るには，別途データを用意する必要があるため，今回は生成の対象外とした．

3.2.1 勝利概要・敗因

(A) 勝利概要と (D) 敗因はテンプレートを用いて生成する．テンプレートは試合中の点差の推移や試合全体の得点，点差などを材料として選択される．

(A) 勝利概要のテンプレートを表 3 に示す．出力内容の＜チーム名＞には「広島が接戦を制した。」のように勝利チームの名前が記述される．例を挙げると，両チームの点差が 2 点差以下で，試合中の最大点差が 3 点以下の場合,「接戦を制した」と出力される．さらにこの時，終了時の得点が 2 点以下だった場合,「投手戦を制した」と出力される．延長戦だった場合，冒頭に「延長戦の末」が追加される．

(D) 敗因のテンプレートを表 4 に示す．表 4 の出力内容の冒頭にはそれぞれ「敗れた＜チーム名＞は、」と敗れたチーム名が記される．表 4 には上から優先度の高い順にテンプレートを示しており，上の方にあるテンプレートほど出力されやすい．

表 3: (A) 勝利概要のテンプレート

条件	出力内容
以下は条件を満たした場合片方のみ冒頭に追加される	
両チーム 6 点以上	「打撃戦の末、」を追加
延長戦	「延長戦の末、」を追加
サヨナラ勝ち	「＜チーム名＞がサヨナラ勝ち。」
点差の条件	
8 点差以上	「＜チーム名＞が大勝。」
4 点差以上，点差 2 倍以上	「＜チーム名＞が快勝。」
2 点差以下で,	
得点 2 点以下	「＜チーム名＞が投手戦を制した。」
最大点差 3 点以下	「＜チーム名＞が接戦を制した。」
逆転の回数	
逆転 2 回以上	「＜チーム名＞がシーソーゲームを制した。」
逆転 1 回	「＜チーム名＞が逆転勝ち。」
全て満たさない	なし

表 4: (D) 敗因のテンプレート

条件	出力内容
逆転された	「リードを守れなかった。」
投手戦に敗れた	「投手を援護できなかった。」
チャンスで凡退	「チャンスを生かせなかった。」
得点 2 点以下，被得点 6 点以上	「投打に振るわなかった。」
得点 2 点以下	「打線が振るわなかった。」
被得点 6 点以上	「投手陣が振るわなかった。」
全て満たさない	「精彩を欠いた。」

表 5: スコア付けの規則

条件	内容
出塁	+0.1
得点	+1
得点語を含む	その得点語に付けられたスコアを加算
「一打」「一発」を含む	凡退だがチャンスの場面であったため，スコアを 0.1 に
犠飛，犠打	+0.0001
それ以外 (凡退)	スコアを 0 に

表 6: ダイジェストの統計

項目	構成要素	平均値	最大値	最小値
文字数	(A) 勝利概要	6.46	31	0
	(B) 打撃	64.71	102	21
	(C) 投手	25.76	46	0
	(D) 敗因	27.50	46	0
	全体	124.5	127	120
単語数	(A) 勝利概要	3.60	15	0
	(B) 打撃	42.24	60	15
	(C) 投手	17.14	39	0
	(D) 敗因	15.53	30	0
	全体	78.52	94	65

3.2.2 打撃

(B) 打撃の生成は以下の手続きで行われる．

まず，各イベントからテンプレートを用いて文を生成する．具体的には，事前に用意した辞書に基づき，イベントからヒットや三振などの結果を表す語を抜き出す．そののちに，抽出した語と，チーム名，イニング，打者名とを組み合わせて各イベントに対応する文を生成する．

次に，生成された文の集合から，重要な文を選択する．文の選択は，各文にスコアを付与し，スコアの高い順に文を選択することで行われる．生成されるダイジェストには文字数の制限があるため，スコアの高い文から順に文字数の制限に違反しない限り文を貪欲に選択することで出力となるダイジェストを構成する文を選択した．

各文に対するスコアの付与は，予め定めた規則と辞書に従って行った．辞書はヒューリスティックに基づいて定めたものである．基本的には辞書に含まれる単語があればスコアを与える規則となっているが，特定の単語が出現する際にはスコアを与えない場合があるなど，細かい条件付けが行なわれている．辞書には得点イベントに含まれる得点語と，それぞれの得点語のスコアが含まれている．例えば,「サヨナラ」は 10,「逆転」は 6 などである．イベントの記述が「一打逆転のチャンスで凡退」などの場合,「逆転」が含まれているが実際には凡退しているため，高い得点が付けられることは望ましくない．そのため，イベント中に「一打」が出現した時点でスコア付けを中断し,「逆転」のスコアである 6 が付与されないようにしている．

4 実験

4.1 データ

実験のため，Yahoo! JAPAN スポーツナビ [1] から，2015 年シーズンの日本野球機構のセントラル・リーグおよびパシフィック・リーグの各試合のイベントと，比較対象としてダイジェストを収集した．各試合のイベントは 2015 年 6 月 15 日から 10 月 29 日の間に行われた試合分，ダイジェストは 2015 年 3 月 23 日の開幕から 10 月 29 日の全試合分を利用した．

各イベントは，表 1 に示したように，イニング，攻撃側チーム名，打者名，打席内容等で構成されている．打席内容の中にはアウトカウントやランナー，得点などの情報が掲載されていることもある．なお，盗塁や選手交代などで打者名が存在しない場合，打者名の欄は空となる．

ダイジェストは，表 2 に示したように，打撃や投手の情報で構成されている．ダイジェストの文字数，単語数の統計を表 6 に示す．統計には中止された試合を除いた 863 試合分のデータを用いた．実験に使用したデータは，取得した各試合イベントのうち引き分けの試合を除いた 453 試合である．引き分けの場合，(A) 勝利概要と (D) 敗因で用いたテンプレートを使用することができないことに加えて，本手法では (B) 打撃はどちらかのチームを主語として出力するものとなっているが，引き分けの場合両チームのイベントを並列して掲載するべきであるため，生成の対象外とした．

4.2 評価方法

各試合のダイジェストを参照テキストとし，これを生成されたテキストを比較することで評価を行う．比較には ROUGE-1[4] を用いた．

個別に生成したダイジェストの構成要素の品質を評価するため，最終的なダイジェスト同士の比較に加えて，個別の要素同士の比較を行った．

なお，表 6 に示すように，ダイジェスト全体は平均して約 125 文字で記述されている．同様の長さのダイジェストを生成するため，生成の際にはまず (A) 勝利概要，(D) 敗因を先に生成し，そののちにダイジェスト全体の平均長から (A) 勝利概要，(D) 敗因の文字数を引いた文字数を目標に (B) 打撃を生成した．

[1] http://baseball.yahoo.co.jp/npb/

表 7: ROUGE-1 による評価の結果

システム出力	参照テキスト	手法	ROUGE-1
B	ABCD	提案手法	0.359
		CF+KP	0.353
		ランダム	0.329
B	B	提案手法	0.483
		CF+KP	0.448
		ランダム	0.402
AD	AD	テンプレート	0.439
ABD	ABCD	提案手法	0.422
		CF+KP	0.413
		ランダム	0.389
ABD	ABD	提案手法	0.488
		CF+KP	0.467
		ランダム	0.433

4.3 比較手法

提案手法によるダイジェストの品質を評価するため，2 つの比較手法を用意した．これらはいずれも (B) 打撃を生成するためのものであり，(A) 勝利概要および (D) 敗因は提案手法と同様の手法で生成した．

- **CF+KP：**CF[2] を用いて各文を構成する単語の重要度を求め，それらの和を各文の重要度とした．そののちに動的計画ナップサックアルゴリズム (KP) を用いて制限長以内で最良の文の組み合わせを得た．CF の計算には Yahoo! JAPAN スポーツナビから取得した各試合のダイジェストを利用した．文の組み合わせの選択には Shuca [3] を用いた．
- **ランダム：**入力されたイベント集合から生成した文集合を，制限長以内でランダムに選択することでダイジェストを生成した．

4.4 結果と考察

ROUGE-1 による評価結果を表 7 に示す．システム出力と参照テキストの項の ABCD は前述したダイジェストの各部分である．

全体的に，提案手法が他の 2 つの手法よりも良好な結果を示した．提案手法の優位性について考えると，まず，各手法の差分は (B) 打撃のみであり，(A) 勝利概要と (D) 敗因は同じものを利用している．提案手法は貪欲法を，CF+KP は KP を用いてそれぞれ (B) を選択しているので，文の重み付けに同一の手法を利用した場合，KP を用いた CF+KP の方が良い結果が得られると予想できる．しかし実際は提案手法の方が優れた結果を示しているため，提案手法が用いた重み付けの方法が，探索アルゴリズムの差異を克服する程度には優れていたと考えられる．提案手法では「一打逆転のチャンスで凡退」など，得点語が含まれているが実際には凡退した場合にはスコアを付けないなど細かい場合分けを行っており，これが重み付けの優位性に繋がったものと考えられる．そのため，KP を用いる際に CF ではなく提案手法における重み付け手法を利用することで，より高い品質のダイジェストが生成できると期待される．

[2]Collection Frequency (CF) は文書集合全体における単語の出現回数である

[3]https://github.com/hitoshin/shuca

5 おわりに

本稿では，野球の試合を対象とし，試合中に生じたイベントを入力とし，ダイジェストを生成する手法を提案した．

今後の課題としては，(C) 投手の部分の生成がある．今回，(C) 投手は生成の対象としなかったが，実際のダイジェストとして投手の情報が欠けているのは不十分であろう．また，引き分けの試合も対象外としたため，(B) 打撃の生成手法を改善しこれを生成の対象とすることも今後の課題としたい．

参考文献

[1] Regina Barzilay and Mirella Lapata. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pp. 331–338, 2005.

[2] Regina Barzilay and Mirella Lapata. Aggregation via set partitioning for natural language generation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pp. 359–366, June 2006.

[3] Ryuichiro Higashinaka, Rashmi Prasad, and Marilyn A Walker. Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (Coling-ACL)*, pp. 265–272, 2006.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop Text Summarization Branches Out*, pp. 74–81, 2004.

[5] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.