

# 学士特定課題研究発表

## サッカーのプレーデータを用いた テキスト速報の自動生成手法の提案




学籍番号: 20B12036

氏名: 田中智可良

研究室: 市瀬研究室

## 1. 背景と目的

# テキスト速報とは

前半 43分	 <p><b>GOOOOOOOOOAL!</b></p> <p>守田が起点となって攻め上がる。久保が守田からの縦パスを受け、相手のタックルにひるまず、旗手につなぐ。旗手はペナルティエリア手前の中央からパスを出すと、相手にブロックされたこぼれ球に反応した古橋がペナルティエリア中央へ抜け出してチャンスを迎える。冷静に右足を振り抜き、ネットを揺らす</p>
前半 42分	 <p>後方のパス回して右サイドにボールを送り、中山が旗手に縦パスを通す。旗手は中央の古橋を狙って斜めのパスを送るも、相手にカットされてしまう</p>
前半 41分	 <p>旗手が左サイドの相手陣中央からクロスを供給。伊東が反応して右から中央へ走り込むが、うまくシュートに持ち込めない</p>

<https://www.jleague.jp/match/japan/2023/101701/livetxt/>

## 概要

- リアルタイムのスポーツの試合内容についてのテキスト
- スポーツサイトに逐次的に更新
- データ会社が手入力

## 用途

- 試合を見ることはできないが流れを追いたいとき
- 試合の流れを文章で振り返りたいとき

## 1. 背景と目的

## サッカーを選んだ動機と研究目的

## 他のスポーツのテキスト速報との比較

試合説明の複雑性	高	サッカー	日本 鈴木が右前方へ長いフィードを蹴り込むと、駆け上がった毎熊がインステップでうまく収める。しかし、すぐさま相手に囲い込まれ、味方へのパスはかなわない
		バスケ	佐賀 0-3 #25 角田 3Pシュート○(3点) ジャンプショット ポインツオフターンオーバー
	低	野球	阪神 近本 2アウト2塁からライトへのヒットで出塁 1,3塁

## 既存研究について

- 野球やバスケットボールのスタッツ表から試合の要約文を生成するベンチマーク(MLB, RotoWire, SportSett)がよくつかわれるがサッカーはほとんどない
- テキスト速報のように、目立ったシーンの説明文を生成するシステムの研究は少ない

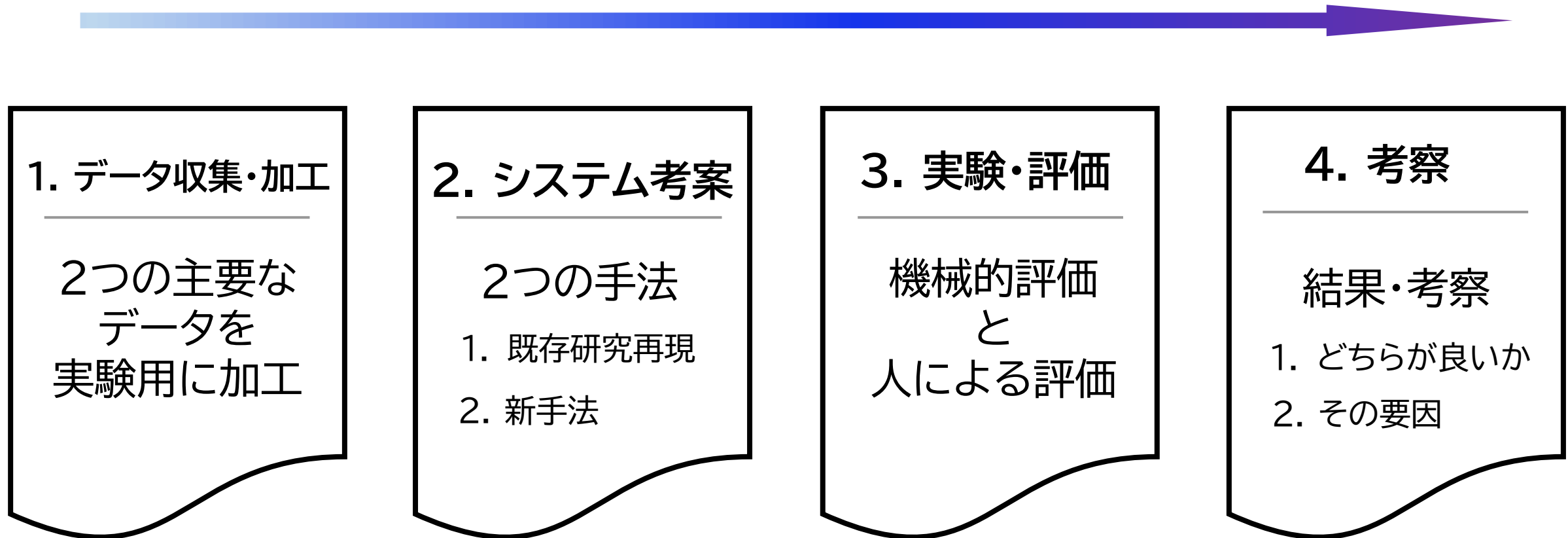
サッカーテキスト速報需要大

試合説明複雑で手入力が大変

AIによる自動生成で省力化

目的

# 研究の具体的な流れ



# データの収集・加工

2021年のJ1リーグの40試合分に関する2つのデータ

## I. プレーデータ

- 1試合につきおよそ2000行のイベントが時系列で記述
- 各イベントに, 選手名, 選手の所属するチーム名, ポジション, フィールド上の位置, アクション内容(パス, シュート, ファウルなど)

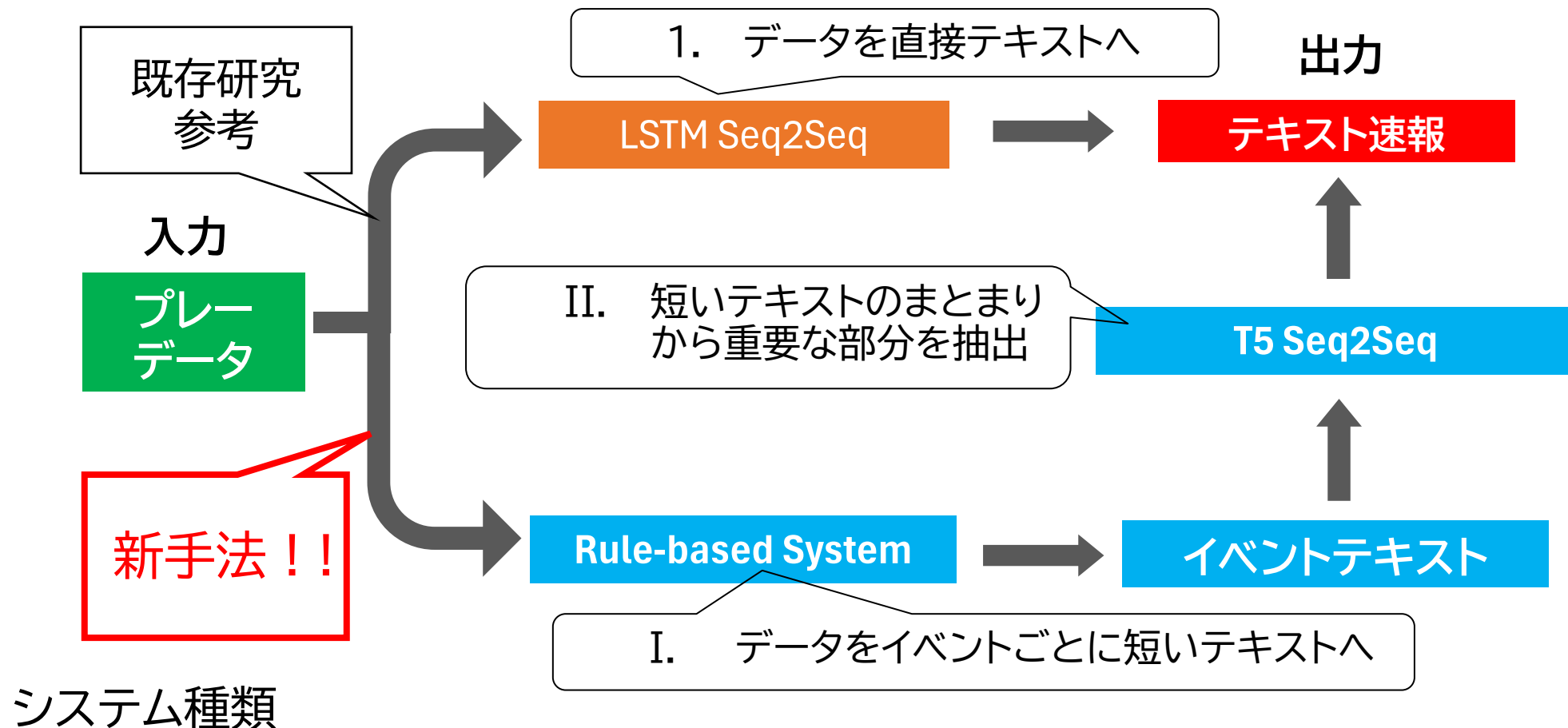
## II. テキスト速報データ

- 1試合につきおよそ70個のコメント
- 目立ったイベントに関するコメント(決定機, ファウル, 交代, VARなど)

<目的達成のための手段>

プレーデータを**入力データ**として, テキスト速報の自動生成を目指す

# 提案システム紹介



NNベース : NN-MLP (Baseline), NN-MHA (MultiHead Attention)

ルールベース : Rule-T5base (param:220M), Rule-T5large (param:770M)

## 2. 研究の流れ

# 評価手法

### 機械的評価

ROUGE : 正解文の中の単語とどれだけ一致している単語が生成文の中にあるか?

#### ① ROUGE-1

- 1単語ごと
- **コンテンツ選択**の精度を評価

入力データのどの部分を使うべきか

#### ② ROUGE-2

- 2単語ごと
- より厳しく**コンテンツ選択**や**文法理解**を評価

### 人手による評価

#### 【意味的な正しさ】

- ① 十分説明できている : 正解文と同義
- ② 最低限説明できている : 矛盾はなし
- ③ 一部できていない : 一部矛盾あり
- ④ 全くできていない : 矛盾しかない

#### 【選手名出力の正しさ】

- ① 十分できている : 位置 and 名前が適切
- ② 一部できていない : 位置 or 名前が不適切
- ③ 全くできていない : 全て違う

# 機械的評価の結果

		ROUGE-1	ROUGE-2
NN-	MLP (Baseline)	0.467	0.202
	MHA	0.442	0.172
Rule-	T5 base	0.476	0.336
	T5 large	<u>0.510</u>	<u>0.376</u>

1. T5 large が最も精度が良かった  
理由：事前学習済みモデル使用が貢献

2. NNベースの方が ROUGE-1 → ROUGE-2への減少 大  
理由：直接数値データ→テキストだと特徴の伝達が難しい



## 2. 研究の流れ

## 人手による評価の結果

意味的な 正しさ	NN- MLP	NN- MHA	Rule- T5large
十分 or 最低限 できている	10 (13%)	14 (18%)	<u>52 (68%)</u>
一部できていない	32 (42%)	42 (55%)	<u>8 (11%)</u>
全くできていない	34 (45%)	20 (27%)	<u>17 (21%)</u>
Total	76 (100%)	76 (100%)	76 (100%)

1. Rule-T5largeの精度が最も良い

2. NNでは MHA > MLP

理由: MHAは正解文の語彙を使わず  
に意味的に正しい文を生成する傾向

選手名出力の 正しさ	NN- MLP	NN- MHA	Rule- T5large
十分できている	11 (14%)	13 (17%)	<u>48 (63%)</u>
一部できていない	21 (28%)	34 (45%)	<u>17 (22%)</u>
全くできていない	44 (58%)	29 (38%)	<u>11 (15%)</u>
Total	76 (100%)	76 (100%)	76 (100%)

1. Rule-T5largeの精度が最も良い

2. NNでは MHA > MLP

原因: NNベースは複数名出力が苦手  
ルールベースは対応可能

## 2. 研究の流れ

# 実際に生成した文の比較例

### 実際のテキスト速報

鳥栖 左サイドからのボールを山下がペナルティエリア手前の中央で収める。  
落としたボールを仙頭がダイレクトで狙うが南にセーブされてしまう

### NNベースによる生成文

鳥栖 小屋松がペナルティエリアから浮き球のパスを供給。受けた小屋松はペナルティエリア手前の中央から左足を振りぬくがDFに当たってコースが変わってしまう。

### ルールベースによる生成文

鳥栖 山下からのパスを受けた仙頭がペナルティエリア手前の中央から右足でシュートを放つ。しかし、南のファインセーブに阻まれてしまう。

### 3. まとめ

# 貢献と今後の課題

## 本手法の貢献

新手法である**ルールベースシステム**のテキスト速報自動生成タスクへの有効性

- 事前学習済みモデルの利用
- 幻覚(入力にない情報を出力してしまう現象)の少なさ
- 固有名詞出力精度の高さ

## 「プレーデータからテキスト速報生成」の欠点

1. すべてのタイプのコメントに対応不可  
— 交代や審判のVAR確認など
2. プレーデータの即時作成

## 今後の課題

- 情報を豊富に含んでいる**映像**からテキスト速報をキャプション
- ルールベース手法を織り交ぜる