実況発話ラベル予測モデルにおける状況認識素性の活用

上田佳祐 [1,2]　石垣達也 [1]　小林一郎 [1,3]　宮尾祐介 [1,2]　高村大也 [1]
[1] 産業技術総合研究所　[2] 東京大学　[3] お茶の水女子大学
{ueda.keisuke, ishigaki.tatsuya, takamura.hiroya}@aist.go.jp
koba@is.ocha.ac.jp　yusuke@is.s.u-tokyo.ac.jp

概要

本稿では実況発話ラベル予測問題として，とくにレーシングゲーム実況に対する発話ラベル予測問題に着目し，レース状況やレーシングカーの状況を認識する新たな素性を用いた予測モデルを提案する．既存研究で取り組まれている 1) 実況テキストが与えられテキストに対する発話ラベルを予測する設定，2) 実況テキストは与えられず対象時刻の発話ラベルを予測する設定の 2 つの問題を扱う．レーシングゲーム実況において実況者は，“プレイヤー第 1 コーナー抜けて一気に加速” といった実況発話を行う．この発話では，プレイヤーの運転する車という言及対象に対し，一気に加速するという動作に関する言及がペアとして述べられている．ラベル予測問題を正しく解くためには，カーブを抜けるというレースの状況や，レーシングカーの速度といった車両に関する状況を正しく認識することが重要である．そこで，本研究では，レース状況や，速度などの車両の状況に関する状況認識素性を発話ラベル予測に活用する．実験より状況認識素性が特に後者のタスクに有効であることが分かった．

1　はじめに

多くのスポーツやゲーム映像には表 1 に示すような実況が付与される．実況者はレースの進捗やイベントを正しく認識し，適切な言及対象や発話内容をリアルタイムに決定した上で，“ターン 1，しっかりブレーキ” といった発話を行う．視聴者は映像を視聴しながら実況を聞くことで，よりイベントを理解し，映像を楽しむことが出来る．

従来，サッカー，野球，レーシングゲーム等を主な対象とし，実況テキストを自動生成する研究が行われている．一方，上田ら [1] は特にレーシングゲームの実況テキストの言及対象や内容を表現した発話ラベルを予測する問題に着目し，1) 実況テキストが与えられ，発話ラベルを予測する問題 (**対象発話ラベル予測**)，2) 過去の発話と次の発話タイミングが与えられ，発話ラベルを予測する問題 (**未来発話ラベル予測**) を提案している．前者は発話テキストに対する言語理解問題であり，実況テキストの要約といった下位タスクへの応用を想定している．後者は未来の発話内容を予測するプランニング問題と捉えられ，実況自動生成への応用を想定している．

時刻	発話テキスト (空撮視点)	ラベル
00:01	本日は長いストレートが有名なラグナセカのレースです.	サーキット/特徴
00:10	プレイヤーは青のポルシェ.	プレイヤーの車/特徴
00:13	ターン 1，しっかりブレーキ.	すべての車/動き
時刻	発話テキスト (ドライバー視点)	ラベル
00:01	今回は 5 番手からスタート.	プレイヤーの車/相対位置
00:10	華麗なスタートを決めていきたい.	プレイヤーの車/未来のプレー
00:13	あぁ追いつけない..	プレイヤーの車/過去のプレー

表 1　レーシング実況発話とその発話ラベルの例．上から 3 発話は空撮映像に対し実況者が付与した実況，末尾 3 発話はドライバー視点の映像に対しプレイヤー自身が付与した実況である．

本研究では従来モデルに対し，レース状況を認識するための素性を追加し，その効果を検証する．既存発話ラベル予測モデルでは，2 つの問題に対する共通の入力素性として，直前の発話テキストおよび発話ラベル，実況者が視聴した映像カメラの視点ラベル，レース開始からの時間経過時間ラベルを入力として用いる．一方，実際の実況においては，“ターン 1、しっかりブレーキ” といった発話のようにレーシングカーの動きに関する言及が多く見られ

る．特に未来発話ラベル予測においては，レーシングカーの速度やブレーキの踏み込みといったレース状況を捉える素性はラベル予測の重要な手がかりとなる．そこで本研究では従来用いられてきた素性に加え，レース状況やレーシングカーに関する状況を捉える素性を用いる．具体的には，実況対象のレーシングカーの速度，ハンドル角度，ラップタイムといった車両の状況に関する素性および周回数，0 から 1 の区間で表現されたレースの進捗状況などレースの進行状況を認識する素性を提案する．実験より提案素性を用いた予測モデルは特に対象発話ラベル予測において，性能を向上させることが分かった．

2 関連研究

実況テキストを対象とした研究は，主に言語生成の分野において，サッカー [2, 3, 4]，野球 [5]，ゲーム映像 [6] などを対象に行われている．一方，実況発話のラベル予測の研究は，レーシングゲーム映像に対する実況データ [6] を対象としたもののみである．

本稿における対象発話ラベル予測は，発話テキストに対するラベル予測であり，従来，電話応答発話 [7] やメール [8] を対象とした言語理解タスクとして取り組まれている．従来，「質問」「意見」といった言語行為論 [9, 10] に基づくラベル設計を用いるのが一般的であるが，レーシングゲームに対してはそのまま適用できない．レーシングゲーム実況に適用可能な発話ラベルとしては，Ishigaki ら [6] が実況テキストの特徴分析に用いたラベルアノテーションを用いる．一方，未来発話ラベル予測タスクは，「次に何について言及するか」を決定する従来の言語生成研究におけるプランニング [11, 12] の問題と捉えることができ，将来的には言語生成問題への応用を想定している．

発話ラベル予測問題は，分類問題もしくは系列タグ付け問題として定式化され，近年はニューラルネットワークによる手法 [1] が用いられる．入力の観点からは多くはテキストのみを扱う手法 [13] であるが，映像や発話音声を用いる手法 [14] が存在する．本研究では，上田らの提案したマルチモーダルなエンコーダを拡張し，レーシングカーやレースに関する状況を認識するための素性を導入する．

3 発話ラベル予測

本節では対象発話ラベル予測および未来発話ラベル予測の問題の入力と出力について定義する．その後，提案モデルについて述べる．

レース状況素性	例
車がピットに入っているか	false
車がピットレーンに入っているか	false
現在ラップ何周目か [0..]	1
現在のラップの進捗 [0, 1]	0.837429
現在のラップが無効かどうか	false
ゲームシステムによる走行の採点	0.0
速度 (km/h)	120.982589
ハンドル角度 (°)	72.911537
座標 (x, y, z) (m)	(5.372770, -10.708132, 843.804809)
道路中央からの位置 (左端=−1, 右端=1)	-0.515301
理論上の最適なコース取りとの差 (m)	0.854022

表 2 レース状況素性の一覧．最後の 2 つのパラメータについては Ishigaki ら [6] によって計算されたものである．

3.1 入力と出力

対象発話ラベル予測問題はラベル付与対象とするテキストを主な入力として扱う．上田ら [1] はさらに，補助的な入力として，直前の発話テキストおよび発話ラベル，カメラの視点情報として「ヘリコプター視点」もしくは「ドライバー視点」のいずれかを取る視点ラベル，さらにレースのスタートから終了までの時間を 4 分割し，それぞれ序盤，中盤の前半および後半，終盤としたレース進捗に関する離散ラベルも用いた．レース序盤ではサーキットの特徴について述べられやすく，終盤ではラップタイムについて述べられやすい．よって，このような離散ラベルはラベル予測問題に対して効果的であると考える．本研究では従来の提案素性に加え，レーシングゲーム Assetto Corsa の API を用いて表 2 に示す情報を用いる．これらのデータは二値もしくは連続値で表現される．例えば，現在のラップの進捗は 0.0 から 1.0 までの間の連続値を取り，ピットに入っているか否かの情報は二値を取る．これらをレースゲーム映像の録画開始時刻から 1 秒ごとに取得し，ラベル分類時刻の過去 10 秒分の数値列をレース状況を捉える素性として用いる．未来発話ラベル予測問題の入力は上述した素性から対象発話テキストを除いたものとなる．

いずれのモデルも既存研究 [6] で定義されたラベルを出力する．ラベル定義を表 3 に示す．実況発話においては，サーキットやレーシングカーなどの言及対象および，タイムやレーシングカーの具体的な動きといったより詳しい内容を含めることが重要である．よって，ラベルは言及対象について表現する

対象サブラベル	例
プレイヤーの車	“ここは華麗な追い抜き，決める”
他の車	“後ろの車に抜かれましたね．”
すべての車	“全車今一斉にスタート．”
サーキット	“ラグナセカは長い ストレートで有名です．”

内容サブラベル	例
相対位置	“プレイヤー現在 2 位．”
絶対位置	“青い車今第 2 カーブに差し掛かり 他の車もそれに続く．”
タイム	“プレイヤー今ゴールラインを超えて 3 分 15 秒．”
直前のイベント	“このミスはタイムに響くぞ．”
将来のイベント	“ターン 15，超えていけるか？”
動き	“プレイヤー，長いストレートで” 追い越していく．
安定したレース	“全車問題なく，順調です．”
特徴	“今回，全車ポルシェ・マカン．”
挨拶	“はい，では実況を始めていきます．”
反応	“おぉ!”

表 3　サブラベルの一覧 [6]．ラベルは 2 つのサブラベルの組み合わせとして定義される．

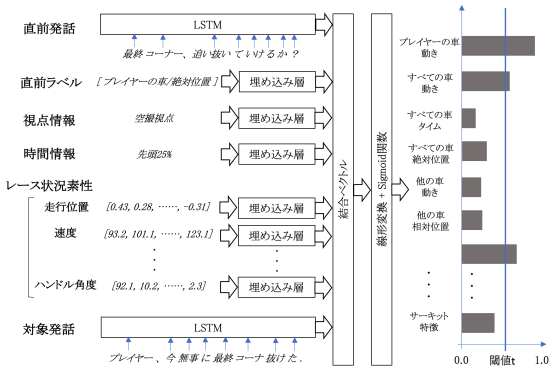

図 1　対象発話を与える設定の予測モデル．次の発話ラベルを予測する設定では，このモデルから「対象発話」が除去される．

サブラベルおよび具体的な内容を定義したサブラベルの 2 種類のペアを 1 つのラベルとして表現し，合計 40 ラベルが定義される．例えば，“ここは華麗な追い抜き，決めていく．” という発話であれば，“プレイヤーの車/動き” というラベル出力が正解となる．1 つの発話に複数のラベルが付与されることを許容するマルチラベル分類の設定となる．

3.2　提案予測モデル

次に図 1 に示す提案ラベル予測モデルについて述べる．このモデルは，上田ら [1] の手法に提案素性を追加したモデルである．2 つの問題設定においてほぼ同様のエンコーダで入力ラベルやテキストを読み込む．ただし，未来発話ラベル予測タスクでは対象発話テキストを入力として与えない．以後，対象発話テキスト予測モデルについて述べる．

まず，テキストデータは単語列で表現する．すなわち，対象発話テキスト $T = \{w_1, .., w_u\}$，直前の発話テキスト $T' = \{w'_1, .., w'_{u'}\}$ となる．なお，対象発話がレース全体の最初の発話である場合は，直前の発話ラベルは与えられない．

T は，MeCab により分かち書きをした後に，各トークンに対応する埋め込み表現へと変換され，単語ベースの双方向 LSTM [15] で読み込む: $\overrightarrow{\mathbf{h}_i} = \mathtt{LSTM}(\overrightarrow{\mathbf{h}_{i-1}}, \mathtt{emb}(w_i))$ および $\overleftarrow{\mathbf{h}_i} = \mathtt{LSTM}(\overleftarrow{\mathbf{h}_{i+1}}, \mathtt{emb}(w_i))$．ここで，emb は対象の単語の埋め込み表現を返す関数である．さらに，ベクトル $\overrightarrow{\mathbf{h}_i}, \overleftarrow{\mathbf{h}_i}$ を結合して，ベクトル $\mathbf{h}_i$ を得る: $\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$．合計 u 個の縦ベクトル $\mathbf{h}_i^T$ を横方向に並べて，行列 $\mathbf{H}$ を得る: $\mathbf{H} = [\mathbf{h}_1^T \cdots \mathbf{h}_u^T]$．直前の発話テキスト T' についても別の双方向 LSTM で同様に読み込み，行列 $\mathbf{H}'$ を得る．LSTM の隠れ状態の行列 $\mathbf{H}$ は，次に深さ r の自己注意機構 [16] に渡される:

$$\mathbf{A} = \mathtt{Softmax}(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{H})), \quad (1)$$

$$\mathbf{M} = \mathbf{A}\mathbf{H}^T. \quad (2)$$

ここで，$\mathbf{A}$ は自己注意機構の重み行列であり，行列 $\mathbf{W}_1, \mathbf{W}_2$ を用いて推定される．$\mathbf{W}_1, \mathbf{W}_2$ のサイズは，LSTM の隠れ状態の次元数を d_{lstm}，自己注意機構の深さを r として，それぞれ $d_a \times 2d_{lstm}, r \times d_a$ となる．ここで，d_a は自己注意機構の重みを推定するニューラルネットワークの行列のサイズを制御するパラメータである．自己注意機構の出力である行列 $\mathbf{M}$ のサイズは $r \times 2d_{lstm}$ となる．行列 $\mathbf{M}$ の各行要素は以下のように結合ベクトル $\mathbf{t}$ に変換される: $\mathbf{t} = [\mathbf{m}_1; ...; \mathbf{m}_r]$．ここで，$\mathbf{m}_i$ は行列 $\mathbf{M}$ の第 i 行目成分を表す．このようにして得られた $\mathbf{t}$ を対象発話テキスト T に対する最終的なエンコーダ出力とする．直前の発話テキスト T' に対しても同様に別の自己注意機構で読み込み，エンコーダ出力 $\mathbf{t}'$ を得る．

構造化データについては，既存研究 [1] の提案素性である視点ラベル D_1，時間ラベル D_2，直前の発話ラベル D_3 に加え本研究ではレーシングカーおよびレース状況に関する状況認識素性を数値列データ $E_1, ..., E_e$ も用いる．これらはそれぞれ別のエンコーダによって分散表現に変換し，結合ベクトルをエンコーダの最終出力とする．視点情報データ D_1 は実況者の視聴した映像に関する情報を表現し，“空撮視点”，“ドライバー視点” のいずれかを取り，時間情報データ D_2 はレース開始から終了までの時間を 4 分割した区間のうち，分類対象発話の発話時刻の

区間を表す．本研究で追加する状況認識素性は，分類対象発話の発話時刻から過去 10 秒分を 1 秒ごとに取得し大きさ 11 のベクトルとして表現する．

既存素性 D_n および提案素性 E_l は，one-hot 表現 $\mathbf{d}_n$ および e_l とし，埋め込み表現に変換する: $\mathbf{d}'_n = \mathbf{d}_n\mathbf{W}_n + \mathbf{b}_n, \mathbf{e}'_l = \mathbf{e}_l\mathbf{W}_l + \mathbf{b}_l$. そして，$\mathbf{t}, \mathbf{t}', \mathbf{d}'_1, ..., \mathbf{d}'_3$ を結合し，この結合ベクトルのサイズがラベル種類数と同一サイズのベクトルとなるよう重み行列 W_3 で線形変換する:$\mathbf{p} = \mathtt{sigmoid}([\mathbf{t}; \mathbf{t}'; \mathbf{d}'_1; \mathbf{d}'_2; \mathbf{d}'_3; \mathbf{e}'_1, ...; \mathbf{e}'_L])$. 最終的に sigmoid 関数によって各次元が確率値となるよう正規化され，各ラベルに対するスコアを表現するベクトル $\mathbf{p}$ を得る．

なお，未来発話ラベル予測タスクにおいては，上式から $\mathbf{t}$ が除去されスコアを表現するベクトル $\mathbf{p}$ を得る．スコアが閾値 t 以上のラベルを最終的に出力する．t は開発セットでの F 値が最も高くなるよう調整する．また，上記の予測モデルは二値交差エントロピー損失を最小化するよう学習される．

4 実験

提案素性の効果を確認するため，それらを用いない既存手法 [1] と比較する．また，ヒューリスティックによるベースラインとして，データセットにおいて出現頻度が上位 k 個のラベルをすべての発話に付与する単純な手法とも比較する．データセットにおける k の平均値は 1.44 であるため，$k = 1$ および $k = 2$ の設定をベースラインとして用いる．

実験には Ishigaki ら [6] のアノテーションデータを用い，訓練データを 80%，評価データを 20%とし 5 分割交差検定により評価した．各分割で訓練データとして用いるデータのうち 25%を開発データとしてハイパーパラメータの調整に使用し，残りのデータをモデルの訓練に用いた．視点ラベル，時間ラベル，直前ラベル，状況認識素性それぞれの埋め込み層の行列サイズは，$2 \times 10, 4 \times 10, 24 \times 10, 11 \times 10$ とした．その他の学習に用いたパラメータなどは既存手法 [1] と同様のものを用いた．

5 結果

結果を表 4 に示す．表の上から順にベースライン手法，未来発話ラベル予測タスク，対象発話ラベル予測タスクでの各モデルの評価値を示す．

未来発話ラベル予測タスクにおいては，従来研究で用いられていた素性のみを用いるモデルは F 値において.298 と低い値を示すが，レース状況素性を加えると.307 に向上した．さらに，レース状況素性のみを用いたモデルは.357 と従来素性との組み合わせモデルよりも性能が高い．これは，従来素性の中にはむしろ性能を劣化させる素性が含まれている可能性や学習データの不足の可能性が考えられる．今後，従来素性および提案素性の適切な組み合わせ手法やデータ拡張手法について検討したい．

	適合率	再現率	F 値
ベースライン:			
k-最頻出ラベル ($k = 1$)	.351	.243	.287
k-最頻出ラベル ($k = 2$)	.297	.412	.345
未来発話ラベル予測タスク:			
従来素性 [1]	.213	.514	.298
従来素性 [1] + 提案素性	.243	.435	.307
提案素性のみ	**.257**	**.605**	**.357**
対象発話ラベル予測タスク:			
テキストのみ	**.773**	**.706**	**.735**
従来素性 [1]	.762	.691	.722
従来素性 [1] + 提案素性	.734	.670	.701

表 4 対象発話ラベル予測タスク及び未来発話ラベル予測タスクでの性能評価.

レース状況認識素性の追加により，未来発話ラベルの予測性能が向上した例について述べる．“インをキープしながら曲がっていく。” という発話に対しては “実況対象の車/動き変化あり” という予測ラベルが正解となる．このラベルを正しく予測するためには，プレイヤーの車の走行状況を正しく認識する必要がある．提案素性に含まれるハンドルの角度が与えられたことで，モデルがプレイヤーの車の走行状況を認識でき，この発話についてラベルを正しく予測出来た．

対象発話ラベル予測タスクにおいては，発話テキストのみを用いるモデルが F 値において性能が最も高く，レース状況素性の追加による性能向上は確認できなかった．これは既存研究 [1] においても議論されているが，発話ラベルはその対象の発話テキストに大きく依存しており，他の追加情報は大きな手がかりとならないにも関わらずモデルのパラメータが増大し学習が難しくなるものと考えられる．

6 おわりに

本研究ではレーシング実況を対象とした発話ラベル予測問題について，レース状況やレーシングカーの状況と捉える素性を用いる手法を提案した．実験より，発話テキストを入力として用いない未来発話ラベル予測タスクにおいて，状況認識素性の追加による性能向上を確認した．

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の助成事業 (JPNP20006) の結果得られたものである．産総研の AI 橋渡しクラウド (ABCI) を利用し実験を行った．

参考文献

[1] 上田佳祐, 石垣達也, 小林一郎, 宮尾祐介, 高村大也. 実況における発話ラベル予測. 情報処理学会自然言語処理研究会 2021-NL-251 (1), pp. 1–6, 2021.

[2] Kumiko Tanaka-Ishii, Koiti Hasida, and Itsuki Noda. Reactive content selection in the generation of real-time soccer commentary. In **Proceedings of the 17th International Conference on Computational Linguistics (COLING1998)**, 1998.

[3] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating live soccer-match commentary from play data. **Proceedings of the AAAI Conference on Artificial Intelligence (AAAI2019)**, Vol. 33, No. 1, pp. 7096–7103, 2019.

[4] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating live sports updates from twitter by finding good reporters. In **2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**, Vol. 1, pp. 527–534, 2013.

[5] Byeong Jo Kim and Y. Choi. Automatic baseball commentary generation using deep learning. **Proceedings of the 35th Annual ACM Symposium on Applied Computing**, pp. 1056–1065, 2020.

[6] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Generating racing game commentary from vision, language, and structured data. In **Proceedings of the 14th International Conference on Natural Language Generation (INLG2021)**, pp. 103–113, 2021.

[7] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. **Computational Linguistics**, Vol. 26, No. 3, pp. 339–374, 2000.

[8] Tatsuro Oya and Giuseppe Carenini. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In **Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL2014)**, pp. 133–140, June 2014.

[9] John Langshaw Austin. **How to do things with words**. William James Lectures. Oxford University Press, 1962.

[10] John R. Searle. **Speech Acts: An Essay in the Philosophy of Language**. Cambridge University Press, 1969.

[11] Karen Kukich. Design of a knowledge-based report generator. In **Proceedings of 21st Annual Meeting of the Association for Computational Linguistics (ACL1983)**, pp. 145–150, June 1983.

[12] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In **Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI2019)**, pp. 6908–6915, 2019.

[13] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In **Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING2016)**, pp. 2012–2021, 2016.

[14] Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. Strong and simple baselines for multimodal utterance embeddings. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2019)**, pp. 2599–2609, June 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural Computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[16] Lin Zhouhan, Feng Minwei, dos Santos Cicero Nogueira, Yu Mo, Xiang Bing, Zhou Bowen, and Bengio Yoshua. A structured self-attentive sentence embedding. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2017.