多人数会話における発話タイミング検出のためのムーブの推定

木村 直登　　藤江 真也　　小林 哲則
早稲田大学 基幹理工学部 情報理工学科
{kimu, fujie, koba}@pcl.cs.waseda.ac.jp

1 はじめに

多人数会話にロボットを参加させるにあたり，ロボットが会話の流れを乱さず，適切なタイミングで発話できることを目的として，ムーブのクラスを自動推定することを試みた．

我々は，3 人以上で構成される多人数会話に参加して，他の会話参加者の発話機会を均等化する会話ロボットを作成している [1]．多人数会話では，時折発話の機会を得られずに「置いてけぼり」となる参加者が発生する．この参加者に対して話を振ることで会話への参加を促すことがロボットの目的であるが，進行中の話の流れを無視して話を振ると，全体の会話の調和を乱す．このため，ロボットは，まず，進行中の会話に割り込み，会話の主導権を得た後に「置いてけぼり」の参加者に話を振るという手順を踏む必要がある．ここで重要なことは，ロボットはどのタイミングで会話に割り込むことが許されるか，すなわち，それぞれのタイミングにおいて誰が発話権を持ちうるかを知ることである．

発話権の所在の可能性を知るためには，ロボットは，進行中の対話における各発話の談話構造上の役割を理解しなければならない。発話の談話機能を分類するものとしては，隣接ペアや TRP(Transition Relevance Place) などがあるが，これらはターンを構成単位として採用している．実際の音声会話では，ひとつのターンの間にポーズが置かれたり，ターンの間に相手の相槌が入ったりすることがあり，実時間での認定が難しい．このため，本研究ではターンより小さい単位であるムーブを採用し，そのクラスによって発話権の所在を推定することを試みた．

ムーブのクラスの分類には様々なものが提案されているが，どれも談話構造を記述することを目的としたものであって，目的とする発話権の所在を直接記述するものではない．ここでは，この目的に従って，ムーブのクラスを再整理した．推定は，系列ラベリングの問題と捉え，識別器には CRF を用いた．識別のための素性には，言語情報，話者情報，韻律情報，時間情報など様々な情報を用いた．

本稿では，特徴量の組み合わせにおける認識率の比較について検討した結果について述べる．

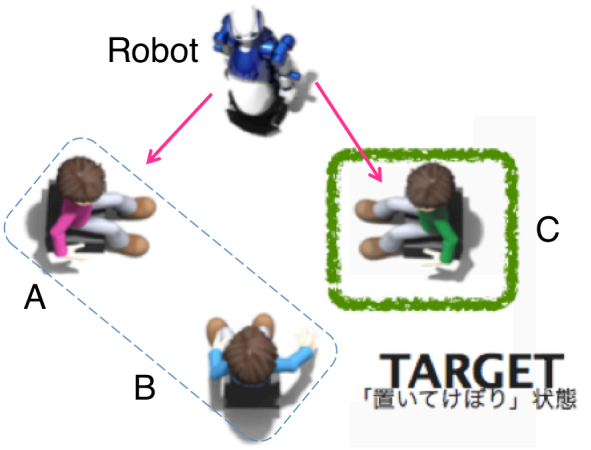

図 1: ロボットは一度会話に参加した後,「置いてけぼり」を救う

2 ムーブ

2.1 ムーブの概念

ムーブとは，J.Sinclair と M.Coulthard が定義した，質問や応答などの談話機能を担う発話単位のことを指す [2]．会話において，ムーブを単位とする発話の役割は，それに与えられたクラスによって表現される。ムーブのクラスは，まず大きく Initiation(開始)，Response(応答)，Follow-up(補足) の 3 種類に分けられる．Initiation は質問などの働きかけを持つ発話を指す．Response は Initiation に対する応答を指す．Follow-up は情報の認識を認定するような補足を指す．ムーブの特性を以下に示す．

- 同一話者の発話の中に複数の役割の発話がある場合 (例：うん，でどうしたの)，それぞれがムーブの単位となる．
- 発話交換は，Initiation - Response の対，Response - Follow-up の対によって成り立つ．
- Follow-up は 2 つ以上続きうる．

ムーブのクラスはさらに詳細なサブクラスに分類される．サブクラスの定義には様々なものがあるが，文献 [5] では次の様に定義されている．Initiation は，Eliciting(誘出)，Directing(指示)，Informing(情報) に区別される．Eliciting は言語反応の要求を指し，Directing は非言語反応の要求，Informing は情報の提供を示す．Response は，Informing(情報)，Behaving(行動)，Acknowledging(認定) に区別される．Informing は情報を持つ応答，Behaving は Eliciting に対応する非言語行動，Acknowledging は発話の受諾を示す．Follow-up は Acknowledging(認定) のみである．

2.2 ムーブクラスの再整理

既に述べたように，我々の目的は，ロボットが参与する多人数会話における各タイミングにおいて，ロボットが発話することを許されるかどうかを知ることである。しかしながら，従来のクラスの分類は直接これを意図して設計されていない。ここでは，発話権の所在の可能性を知るという我々の目的に沿って，文献 [3] の分類を再整理する．

今回，我々が扱う会話は，非言語行動を想定していない。このため，Initiation における Directing，Response における Behaving は対象にしない．

Initiation における Eliciting と Informing は，発話後の発話権の所在の観点からは区別が必要である。なぜなら，Eliciting では，質問を向けられた受け手が次の発話を行うことを期待されるのに対し，Informing は受け手が決まっていないため，話者を含めた全参加者が次の発話の権利を持つからである。

Response における Acknowledging には注意を要する。例えば,「はい」という発話は，質問・依頼に対する応答の場合（「はい（わかりました)」と，コミュニケーションにおける相槌（「はい（聞いていますよ，続けて下さい)」）の場合が考えられる。質問に対する応答としての Acknowledging では，応答の後誰もが発話権を持ちうるのに対し，相槌としての Acknowledging では，話者が発話を継続することが求められる。このため，両者は区別されるべきである。そこで，相槌としての Acknowledging を Back-Channel として新しく定義する．また，Response における Informing と質問・依頼に対する応答としての Acknowledging は，どちらも話者を含めた全参加者が次の発話の権利を持つため，区別の必要がない。これらは，Response として統合する。

Follow-up はサブクラスを持たないため，そのまま用いる．また，実際の会話では割り込みによって発話が途中で遮られ，談話構造上の役割がなくなることがよくあるが，こうしてできるムーブの断片を上の分類では扱えない。例えば,「えっとねえ」のタイミングで他の参加者に割り込まれた場合,「えっとねえ」に与えるべきムーブのクラスがない．そこで，情報を持たない発話を意味する「Portion」を新しく追加する．以下に，ここで使用するムーブの分類を示す．

Eliciting (誘出) - 例：ハリーポッター見た事ある？

Informing (情報) - 例：一作目が好きなんだ．

Response (応答) - 例：あるよ．

Follow-up (補足) - 例：そうなんだ．

Back-Channel (相槌) - 例：うんうん．

Portion (一部) - 例：えっとねえ，

3 データ

ムーブの推定に用いた発話のデータについて述べる．用いたデータは「千葉大学 3 人会話コーパス」である [4]．内容としては，千葉大学で収録された同性 3 人からなる友人同士 12 グループ (男女 6 組ずつ，当時 18〜33 才) の雑談を収めたものである．各会話で決まったトピックについて自由に会話を行っている．コーパスには，音声データと発話の書き起こしテキストが付与されている．データ数は 9 分 30 秒 × 12 グループである．

しかし，コーパスに収録された書き起こしテキストをそのまま学習に用いることは望ましくない．コーパスの各発話の書き起こしテキストは LUU という発話単位で区切られている．LUU とは Long Utterance Unit の略で，話し手と聞き手が行為や情報を交換する際の基本単位のことを指す [5]．一方でロボットは音声区間検出 (以下 VAD) を用いて発話の認識を行う．LUU で区切られた発話では，実環境と発話の切れ目が異なるため，ロボットに組み込む観点では有効ではない．

そこで，コーパスに収録された音声に対し VAD を適用して音声区間に区切った．その結果，総発話は約 2000 発話であった．この発話に対して，人手で 6 種類のムーブを付与した．表 1 に発話とムーブの例を示す．表におけるムーブはイニシャルで表す (例:Eliciting は「E」)．これらの 12 グループのデータから，1 つを評価データ，それ以外の 11 グループを学習データとし，推定を行った．

表 1: VAD によって区切られた各発話の例

話者	発話	ムーブ
A	えどこに落としたの	E
B	んー多分	P
B	んーと自分ちの近くの駅で改札をくぐって次の改札を出るまでのどこか	R
C	ええ出てこなそう	F
B	あたしでもなんかねポケット入れてたんだよねあほだから	I
A	あー	B
B	半分使って七万	I

表 2: 特徴量の例

形態素	形態素	形態素	形態素	品詞	品詞	品詞	品詞	発話者	韻律クラス	発話長クラス	CRF ラベル
な	年	じゃ	ない	助動詞	名詞	助詞	助動詞	Continue	i1	t6	Eliciting
null	null	null	うん	nullPos	nullPos	nullPos	感動詞	Changed	i4	t1	Response
null	null	null	で	nullPos	nullPos	nullPos	接続詞	Changed	i3	t1	Portion
出	て	くる	とか	動詞	助詞	動詞	助詞	Continue	i4	t6	Portion

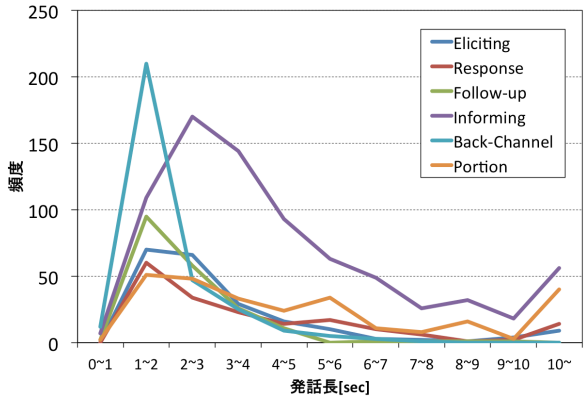

図 2: 発話長とムーブの数の関係

4 識別手法

音声区間の情報をもとに付与された情報を用いて，ラベルの識別を行った．システム動作中に取得可能なものを特徴量とした．具体的には，ロボットが実環境で動くことを想定し，言語情報，話者情報，韻律情報，時間情報である．学習・識別には，系列ラベリング問題を解く際に一般的に用いられている CRF(条件付き確率場) を用いた．

4.1 特徴量

言語情報として，形態素と品詞の情報を利用した．コーパスの書き起こしテキストの各発話に対して，MeCab を利用して形態素解析を行った．形態素の中でも，ムーブの推定に有効なのは発話末の情報であると考え，各発話の最終形態素 4 つと各形態素の品詞情報を特徴量とした．

韻律情報には，岩田が提案した発話末の代表的な F0 パターンを用いた [6]．発話の最終音節の母音区間の F0 をもとに 6 つのクラス (i1〜i6) に分類される．F0 とは基本共振周波数を指す．F0 を用いることで，例えばテキストでは同じ表記でも，質問なのかどうか等を判断できる (例：「そうなの」を，質問である「そうなの？」か共感の「そうなの」かを区別する)．クラスの番号を与える理由は，識別器として利用した CRF では連続値が扱えないため，離散化する必要があったからである．

時間情報として，各発話の発話長ごとの頻度をもとに分類したクラスの番号 (t1〜t6) を特徴量とした．発話の時間ではなくクラスの番号を与えた理由は，CRF が連続値を扱えないからである．図 2 に発話長ごとの頻度を示す．例えば，Back-Channel は 1〜2 秒の発話が多く，Informing は 2〜3 秒の発話が多いことがわかる．この表から，それぞれのムーブの数に特徴のある 6 つ（0〜2 秒，2〜3 秒，3〜5 秒，5〜6 秒，6〜10 秒，10〜秒）の区間をクラスとし，そのクラスの番号 (t1〜t6) を特徴量とした．

話者情報は，話者が変わったかの 2 値 (Continue, Changed) を与えた．これは，ムーブの推定においては，話者が「誰か」(A，B，C) の情報よりも「変わったか」の情報の方が重要であると考えたからである．以上に挙げた特徴量をまとめ，表 2 に示す．

4.2 CRF

学習，識別に用いたのは，CRF(条件付き確率場) である．CRF は，系列ラベリング問題のために設計された識別モデルである．CRF を用いる利点は構造学習をできる点である．素性テンプレートを用いることで，過去の情報を考慮することができる．各発話のムーブの推定では，推定を行う 1 つ前の発話の情報を考慮している．ツールには CRF++を利用した [7]．

5 実験

4.1 で述べた特徴量の組み合わせでムーブの認識実験を行った．組み合わせは，1. 言語情報+話者情報，2. 言語情報+話者情報+韻律情報，3. 言語情報+話者情報+時間情報，4. 言語情報+話者情報+韻律情報+時間情報，である．各組み合わせにおける認識率の結果を表 3 に示す．

表 3: 認識率の比較

言語+話者	言語+話者+韻律	言語+話者+時間	言語+話者+韻律+時間
60.8 %	62.3 %	63.1 %	63.7 %

表 3 から，全ての特徴量を含めた場合が最も性能が高いことがわかった．

表 4: Confusion Matrix

		推定結果					
		Eliciting	Informing	Response	Follow-up	Back-Channel	Portion
正解	Eliciting	134	114	27	3	3	13
	Informing	64	495	51	13	9	48
	Response	22	79	102	7	14	11
	Follow-up	5	36	10	61	18	1
	Back-Channel	0	3	5	7	309	6
	Portion	12	87	24	3	7	133

6 考察

認識率を比較すると，言語情報，話者情報，韻律情報，時間情報の全ての特徴量を含めた場合の精度が一番高いことがわかったが，言語情報と話者情報のみの精度との差が約 3 ポイントと僅差であった．

表 4 に Confusion Matrix を示す．いずれのムーブも正解のムーブを推定した割合が高い結果となったことがわかる．しかし，全ての認識結果は Informing に間違えて推定した数も多くなっている．この原因の 1 つに，倒置法を用いた発話が多かったことが考えられる．例えば,「なんのポイントなのそれは」や「どっちがやったの太田と田中」である．これらの 2 つの発話は Eliciting が正解であるが，Informing と間違えている．今回用いた言語情報は，最後の 4 形態素のみを用いているため，このような場合有効に機能しなかったと考えられる．

次に，各誤りをロボットを動かす際のリスクの観点で考察する．ロボットが，Response，Follow-up の後に話すという行動をとることにすれば，リスクの高い認識誤りは，

- Eliciting，Back-Channel，Portion の発話を Response に間違える
- Eliciting，Back-Channel，Portion の発話を Follow-Up に間違える

であると考えられる．Eliciting，Back-Channel，Portion の後は，会話に入ってはいけないタイミングであるため，間違えるのはリスクが高い．Eliciting，Back-Channel，Portion を Response または Follow-up と間違えているのは総発話のうち約 3.4 % と，多くない．リスクの観点では，良い識別器であると考えられる．

7 まとめと今後の予定

多人数会話において，発話タイミング検出のためのムーブの拡張と推定を試みた．ムーブを拡張し，推定を行った．特徴量の組み合わせの比較を行ったところ，言語情報，話者情報，韻律情報，時間情報の全ての特徴量を含めた場合の認識率が 63.7 % で一番高いことがわかった．今後は，作成した識別器をロボットに組み込み，ロボットが会話に入る際に「ムーブ」を利用することが適切かどうかの評価を行う．また，特徴量の再構成，推定の誤りを考慮したモデルの作成と，実際に発生する音声認識結果のエラーを踏まえたモデルの作成なども検討する．

参考文献

[1] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, Tetsunori Kobayashi. “Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant”, Computer Speech and Language, Vol.33, no.1, pp.1-24, 2015

[2] Sinclair.J, M.Coulthard “Toward an analysis of discourse: the English used by teachers and pupils. ”, Oxford: Oxford University Press., 1975

[3] 石崎雅人, 伝康晴, ”言語と計算 3 談話と対話”, 財団法人東京大学出版会 ,p137-175, 2001

[4] Den, Y. & Enomoto, M. (2007). A scientific approach to conversational informatics Description, analysis, and modeling of human conversation. In Nishida, T. (Ed.), Conversational informatics An engineering approach, pp.307-330. Hoboken, NJ John Wiley & Sons.

[5] Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida, “Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme.” Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010), pp.2103-2110, 2010

[6] Kazuhiko Iwata, Tetsunori Kobayashi, ”Speaker’s Intentions Conveyed to Listeners by Sentence-Final Particles and Their Intonations in Japanese Conversational Speech”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p6895-6899, 2013

[7] 工藤拓, 山本薫, 松本裕治, ”Conditional random fields を用いた日本語形態素解析”, 情処学 NL 研報, p161-163, 2004