The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

マルチモーダル特徴量を用いた談話セグメントの検出

Identifying Discourse Boundaries in Group Discussions using Multimodal Features

冨山 健[*1] Ken Tomiyama　二瓶 芙巳雄[*1] Fumio Nihei　高瀬 裕[*2] Yutaka Takase　中野 有紀子[*2] Yukiko Nakano

[*1] 成蹊大学大学院理工学研究科 Graduate School of Science and Technology, Seikei University

[*2] 成蹊大学理工学部 Faculty of Science and Technology, Seikei University

This study proposes models for detecting conversation boundaries in group discussions. First, we created a multimodal embedding space using an autoencoder, and applied a similarity-based approach to detect the discussion boundary. As the second method, we annotated conversation boundaries and created unimodal CNN models for language, audio, and head motion information. Then, created multimodal models by concatenating the output of unimodal models. In the evaluation experiment, we found that language information was the most useful modality, but by combining with audio and head motion modalities, the CNN-based models more accurately predict the conversation boundaries.

1. はじめに

自然言語処理の分野では，テキストを複数の意味的まとまりに分割することをテキストセグメンテーションという．テキストを分割することは文章の要約や検索などに有益であると考えられる．議論においても，セグメントに分割することができれば，話題構造の認識やファシリテーション支援にも応用できる．例えば，セグメントが短ければ，そのトピックについて議論が十分でない可能性があり，ファシリテータの介入決定の情報となりうる．

文章のセグメンテーション手法は，これまでに多くの研究が行われてきたが，高い精度をもつ手法は未だ確立されておらず，また，対話文に適用した研究は少ない．多くの研究では語彙結束性をセグメント境界の識別モデルの作成などに使用しているが，対話においてはそのスパース性が問題になってくるなど，難しい課題である．

そこで本研究では，議論参加者の言語・非言語情報に基づいた，会話のセグメント境界を検出するモデルを提案する．

2. 関連研究

[1]では語彙結束性をとらえ，文章を分割する手法 TextTiling を提案している．TextTiling は，文章に 2 つの連続する窓をスライドさせ，類似度を算出することで境界を検出する手法である．これを基にした手法も多く提案されていて，広く知られた手法である．また，[2]は新聞や TV ニュースの書き起こしを用いて，手掛かり語彙や前後に出現する語彙情報などから識別関数を作成している．[3]では複数人対話コーパスのテキスト，音声情報を用い，手掛かり語彙，沈黙時間，オーバーラップ時間などを基にしたトピックシフト識別モデルを作成している．

3. 議論コーパス

3.1 議論収集実験

本研究で使用する議論コーパスについて述べる．

1 グループ 3 人 9 グループ，男女 27 人の実験参加者が議論を行った．課題は外国人へ向けた一日観光コースをテーマに 30 分間話し合い，その後 5 分間でアピール点を決定してもらうというものである．最終的に「観光客の国籍と観光コースのキャッチフレーズ」「観光コースの詳細」「アピール点」を記入した用紙を提出させた．議論は「対象の外国人の国籍」「コースのキャッチフレーズ」「観光コースの作成」の順に行うよう指示した．議論前に参加者には議論終了後に作成した観光コースについて「観光スポットの数が多い」などの評価点によって評価されることが伝えられた．

図 1：実験環境

図 1 に実験環境を示す．Kinect による各参加者の追跡データ，頭部に装着したモーションセンサによる 20 ms 毎の加速度・角速度・地磁気のデータ，ヘッドセットマイクによる各参加者の音声データ，ビデオカメラによる議論全体の動画が収集された．また録画した動画から議論の書き起こしを作成した．

3.2 セグメント境界のアノテーション

収集した議論に対して，会話のセグメント境界のアノテーションを行った．初めに，2 名のアノテータが試行と議論を繰り返し，コーディングスキームを作成した．セグメント境界になる発話のルールとして次の 3 つのルールを定めた，1.「話題が移った発話」，2.「会話のレベルが下がる(上がる)発話」，3.「書き込み作業について言及した発話」．1 番目のルールはたとえば，「旅行時期」から「旅行場所」へ話題が移った場合である．2 番目のルールはたとえば「旅行場所」からより詳細な話題「京都」へ移った場合である，3 番目のルールは議論内容に関係するトピックではないが，本研究で使用したコーパスでは多々見られたため，追加した．

連絡先：中野有紀子，成蹊大学理工学部，y.nakano@st.seikei.ac.jp

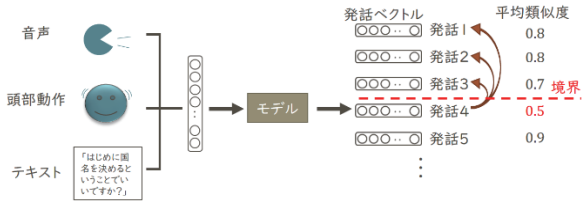

図 2：類似度に基づくセグメンテーション

この作成したコーディングスキームをもとに，全 9 グループについて自動で検出された音声区間に対して，1 名のアノテータがセグメント境界を付与した．ただし，前章で述べた通り，議論 35 分間の内，最後の 5 分間は基本的には用紙への書き込み作業になるため，それを除いた 30 分をアノテーションの範囲とする．その結果，各グループにつき平均約 35 個のセグメント境界が付与された．

4. 発話ベクトルを用いた類似度算出モデル

セグメント内の発話は結束性のある内容を持っていると考えられる．そこで，発話をベクトルで表現できれば，発話間の類似度を算出し，類似度に基づき境界を決定する手法を利用できると考えた．また，議論上で人間は言語だけでなく非言語情報も用いているので，マルチモーダルな情報を用いた議論のセグメント境界の識別を目的とする．そこで，マルチモーダルな情報を基に AutoEncoder を用いて発話を表現するベクトルを作成し，発話ベクトルの類似度を算出し，境界の識別を試みる．提案手法の概略図を図 2 に示す．

4.1 特徴量

セグメント境界の識別に使用する特徴量を下に記す．これらは発話単位で算出される．

(1) 新規/既出名詞の数

発話の書き起こしを形態素解析器 Mecab にかけ，名詞を抽出する．抽出した名詞がそれまでの議論で出現した名詞と一致すれば既出名詞，一致しなければ新規名詞とし，それぞれの個数を数え上げる．一度出現した名詞は既出名詞一覧に登録され，以後の新規/既出名詞の判断に用いられる．

(2) 現発話と直前の発話間の一致/不一致名詞数

直前の発話と現発話に共通する名詞の数を共通名詞数とする．また，現発話に出現し，直前の発話には出現しない名詞の個数を不一致名詞数とする．

(3) 現発話と直前の発話間の一致/不一致動詞数

直前の発話と現発話に共通する動詞の数を共通動詞数とする．また，現発話に出現し，直前の発話には出現しない動詞の個数を不一致動詞数とする．

(4) 発話長（時間長と形態素数）

発話長を示す特徴量として，発話区間の時間長 ms と発話を構成する形態素数の 2 種類を使用した．発話者の話速により，同じ時間長でも形態素数が異なるため，発話の形態素数も発話長として特徴量とした．

(5) オーバーラップ時間

2 つの連続する発話のオーバーラップ時間 ms を特徴量とする．1 人の発話に他の 2 人の発話がオーバーラップしている場合は，それらの合計を算出する．

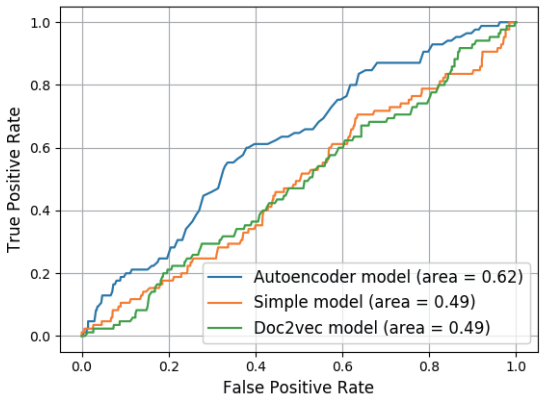

図 3：3 つのモデルの ROC 曲線と AUC による比較

(6) 頭部回転量

発話が行われている時間を対象に，20 ms の時間窓を設定し，各区間における頭部のヨー角回転量 degree をデータとする．各区間における最大値・最小値・平均値・分散値をそれぞれ特徴量とする．

(7) 音声インテンシティ

音声解析ソフト Praat を用いて，収録した発話音声のインテンシティ dB を計測する．各発話区間において 10 ms の時間窓を設定し，その区間におけるインテンシティの最大値・最小値・平均値・分散値をそれぞれ特徴量とする．

(8) 合成加速度

議論参加者の後頭部につけられたモーションセンサの出力から下式を算出することで，合成加速度 HA を得る．ここで，x, y, z はそれぞれ x 軸，y 軸，z 軸方向の加速度の値である．

$$HA = \sqrt{x^2 + y^2 + z^2}$$

発話毎に，算出された合成加速度から最大値，最小値，平均値，分散を計算する．

(9) 合成加速度の Wavelet 変換特徴量

(8)で得られた合成加速度に Daubechies ウェーブレットを用いた多重解像度解析を行い，最下位レベルの信号の最大値，最小値，平均値，分散値を計算する．

(10) Doc2vec 特徴量

Doc2vec [4] を日本語の Wikipedia 記事を用いて学習し，各発話について 200 次元のベクトルを得る．

4.2 識別モデルの作成

発話ベクトル作成に使用する AutoEncoder は入力層，中間層，出力層の 3 層からなり，中間層と出力層でそれぞれ ReLu と Liner の活性化関数を用い，コスト関数は平均二乗誤差を使用する．入力された 241 次元のベクトルは中間層で 150 次元まで圧縮される．これを発話ベクトルとして類似度の算出に用いる．

AutoEncoder の訓練には 9 グループの内 7 グループ（4044 発話）を使用し，残りの 2 グループ（1124 発話）を検証に使用した．1000 epoch で訓練を行ったが 300 epoch にはモデルは収束したため，300 epoch 時点のモデルを用いて各発話に対して 150 次元の発話ベクトルを作成した．

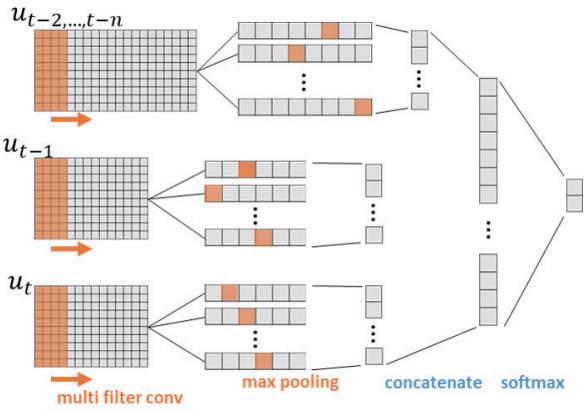

図 4: 言語モダリティモデルのアーキテクチャ

現発話と前 3 発話との発話ベクトルのコサイン類似度をそれぞれ算出し，その平均値を計算する．平均値が任意の閾値を下回れば，その発話を境界と識別する．

4.3 識別モデルの評価と議論

閾値によってセグメント境界かどうかの精度が変化するため，ROC 曲線と AUC で評価する．また，あいづちやフィラーなどの意味のない発話を考慮し，発話長が 0.8 秒以内の発話は除いた．比較のため，提案手法である AutoEncoder で特徴量ベクトルを圧縮したベクトルを用いるモデル（AutoEncoder model）の他に，圧縮せず特徴量を連結させただけのベクトルを用いるモデル（Simple model），Doc2vec 表現だけベクトルを用いるモデル（Doc2vec model）の 3 つを用意し，同様に評価した．その結果，ROC 曲線と AUC は図 3 のようになった．AUC は AutoEncoder model が 0.62, Simple model は 0.49，Doc2vec model は 0.49 となり，AutoEncoder model が一番性能が良い結果となった．

5. 教師あり学習によるセグメント境界識別モデル

各発話について境界であるかどうかを識別する 2 値分類モデルを作成する．教師データとして 3.2 節で作成したアノテーションを用いる．言語情報として発話の書き起こしテキスト，非言語情報として頭部動作，音声を用い，それぞれの 3 つのユニモーダルモデルを作成する． 言語モダリティモデルは Word2vec から得た各発話内の単語ベクトルを入力とし，頭部動作と音声モーダルモデルはそれぞれスペクトログラムを入力とするモデルを作成する．

5.1 入力データの作成

言語モダリティモデルの入力として，Word2vec から得た発話内の単語ベクトルを単語数分連結させた（200 次元×単語数）のサイズの 2 次元配列を発話毎に作成した．

頭部動作スペクトログラムは議論参加者毎に作成する．頭部に装着したモーションセンサから得た角速度から合成角速度を算出し，フーリエ変換することで 25 の周波数解像度を持つスペクトログラムを作成した．音声モーダルモデルの入力となる音声スペクトログラムも同様に議論参加者毎に作成する．ヘッドセットマイクより収録した音声から 32 の周波数分解能を持つスペクトログラムを作成した．また，各データに対して標準化を行った．

さらに，全議論中の最大発話長でゼロパディングを行う．その結果，それぞれの入力サイズは言語モダリティモデルが 217(コ)×200(次元)，頭部動作モデルが 761×25(周波数分解能)，音声モデルは 761×32(周波数分解能)となった．

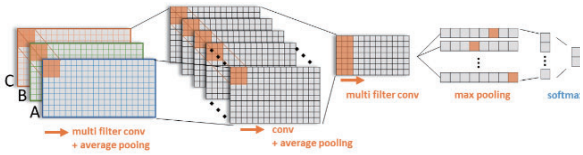

図 5: 頭部動作と音声モダリティのためのアーキテクチャ

5.2 言語モダリティモデルのアーキテクチャ

[5]を参考にアーキテクチャを作成した．アーキテクチャ図を図 4 に示す．現在の発話 u_t(t：発話数)，一つ前の発話 u_{t-1}，過去の発話 $u_{t-2,\ldots,t-n}$(n: 履歴発話数)の 3 つの入力を持つ．各入力されたベクトルは中間層で時間軸方向の 1D-convolution と global max pooling 後に 3 つの出力を連結させて，最終的に softmax 層で境界かどうかの予測が行われる．訓練時のカーネルサイズ，フィルタ数を表 1 に示す．

5.3 頭部動作と音声モーダルモデルのアーキテクチャ

頭部動作と音声モーダルモデルは入力がスペクトログラムであるため，同等のアーキテクチャを用いた．3 人の議論参加者のスペクトログラムを 3 チャンネルとして入力する．参与役割を考慮して，スペクトログラムの順序は優位性の近似値である現発話までの累積発話数[6]で並び替える．入力されたベクトルは中間層で複数の 2D-covolution 層と average pooling の後，時系列方向の 1D-convolution 層と global max pooling を通り，softmax 層で境界かどうかの予測を行う．訓練時には頭部動作モダリティ，音声モダリティともにそれぞれ 2 層の 2D-convolution と average pooling 層を用いた．訓練時のカーネルサイズやフィルタ数を表 1 に示す．

表 1: 各モダリティモデルとカーネルサイズ

	2D-conv	1D-conv	2D-average pooling	2D-conv フィルタ数	1D-conv フィルタ数
テキスト		4			32
頭部動作	(2, 2)	8	(2, 1)	8	32
音声	(2, 2)	8	(2, 1)	8	32

5.4 モデルの訓練と評価

言語，頭部動作，音声のユニモーダルモデルを作成し，3 つの組み合わせも考慮して合計 7 つのモデルを作成する．マルチモーダルモデルは各ユニモーダルモデルの全結合層の入力を連結し softmax 層への入力とすることでモデルの融合を行った．また，モデルの重みの初期値としてユニモーダルで学習した重みを使用する．言語モダリティモデルでは履歴数 10 発話し過去の発話は 8 発話でモデルを作成する．頭部動作モデルは各 convolution 層では10^{-7}の L2 正則化に加え，pooling 層後に 0.25 の dropout を行う．音声モデルも同様に各 convolution 層で10^{-8}の L2 正則化に加え，pooling 層後に 0.25 の dropout を行った．

作成されたモデルは各発話に対してセグメント境界であるかどうかの予測値を出力する．予測値が 0.5 を下回る，つまり境界であると判別された発話が少なかったため，ROC とその面積である AUC でモデル性能の評価を行う．

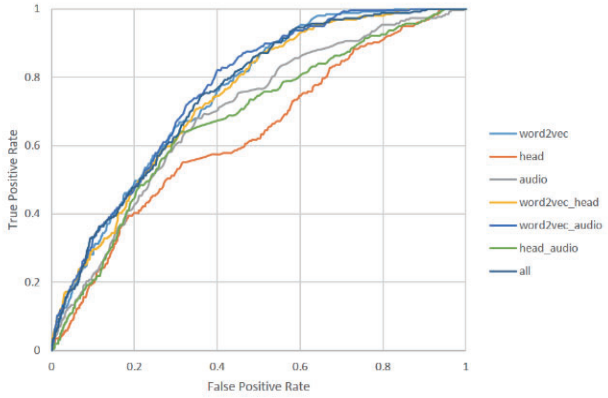

図 6: cross validation での ROC 曲線と AUC

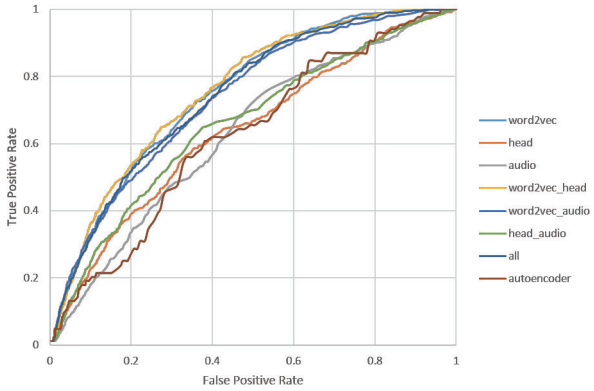

図 7: leave one Group out での ROC 曲線と AUC

全データを 5 分割し，そのうち 1 つを評価データとし，残りのデータで cross validation を行った．5 分割すべてに全グループのデータの発話が含まれるように，グループごとにセグメント境界の発話とそうでない発話に分割し，5 分割したものを統合した．

具体的な訓練・検証・評価データのサイズはそれぞれ約 2500 発話，約 830 発話，約 827 発話である．言語，頭部動作，音声モダリティのモデルをそれぞれ w2v, head, audio とする．これに加え，2 つのモダリティを組み合わせたモデル，言語＋頭部動作(w2v_head)，言語＋音声(w2v_audio)，頭部動作＋音声(head_audio)，全てのモダリティを統合したモデル(all)の 4 つのマルチモーダルモデルを作成した．図 6 にモデル毎の 4 hold の平均 ROC 曲線と平均 AUC を示す．w2v_audio, w2v, all の順に性能が高くなっていることから，言語情報に音声情報を追加すると性能が改善されることがわかる．

しかし，cross validation の評価は訓練・検証・評価で同一グループ内の発話が含まれてしまっている．そこで AutoEncoder モデルとの比較のため，グループ 2 と 7 の 2 グループを評価データとして，残りの 7 グループで leave one Group out を行った．cross validation と同様に各 7 個の検証データごとに作成されたモデルの平均 ROC と AUC を算出した結果を図 7 に示す．w2v_head, w2v, all の順に性能が高く，cross validation での評価とは異なり，言語情報に頭部動作情報を追加したモデルが一番良い性能であった．これは頭部情報がグループによる違いが少ない，識別に有効な情報を持っている，あるいは音声情報がグループ毎の性質に依存する情報である可能性がある．また，音声と頭部動作のモデルは僅差ではあるが，すべての教師あり学習モデルが AutoEncoder モデルの性能を上回った．

また，f-measure が最大となる時の閾値を用い，セグメント境界であると判定する precision, recall, f-measure を算出した．その結果を表 2 に示す．f-measure に着目すると，w2v_head モデルが最もよい性能であった．

表 2: モデル性能の比較

	w2v	head	audio	w2v_head	w2v_audio	head_audio	all	AutoEncoder
平均AUC	0.747	0.634	0.626	0.753	0.732	0.655	0.741	0.624
precision	0.211	0.177	0.131	0.253	0.228	0.181	0.223	0.148
recall	0.549	0.357	0.731	0.440	0.423	0.405	0.497	0.560
f-measure	0.305	0.237	0.222	0.321	0.297	0.25	0.308	0.234

6. まとめ

本研究では，言語・非言語情報に基づく，会話のセグメント境界の検出モデルを提案した．言語情報は最も有用なモダリティであったが，音声や頭部動作と組み合わせることで，より精度よく会話セグメント境界を検出できることが分かった．本研究では，議論参加者の身体情報として頭部動作情報を入力としたが，実際の議論ではポーズやジェスチャーも重要な意味を持っている．特にポーズは，例えば，前のめりになれば議論に対する積極性を示すなどセグメント境界の識別に有効であると考えられる．よって，頭部動作だけでなく姿勢位置などの全体的な身体情報も有効になると考えられる．

謝辞：本研究は，科学技術振興機構(JST) 戦略的創造研究推進事業(CREST)「実践知能アプリケーション構築フレームワーク PRINTEPS の開発と社会実践」(JPMJCR14E3)，および理化学研究所革新知能統合研究センターの支援を受けたものである

参考文献

[1]. Hearst, M.A., Multi-paragraph segmentation of expository text, in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 9-16, 1994.

[2]. Beeferman, D., A. Berger, and J. Lafferty, *Statistical Models for Text Segmentation.* Machine Learning, 1999. **34**(1-3): p. 177-210.

[3]. Galley, M., et al., Discourse segmentation of multi-party conversation, in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. pp. 562-569 2003.

[4]. Le, Q. and T. Mikolov, Distributed representations of sentences and documents, in *Proceedings of the 31st International Conference on Machine Learning - Volume 32*, JMLR.org: Beijing, China. p. II-1188-II-1196, 2014.

[5]. Seokhwan Kim, Rafael E. Banchs, Haizhou Li, Exploring Convolutional and Recurrent Neural Networks in Sequential Labelling for Dialogue Topic Tracking, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 963–973, 2016.

[6]. 吉野 亮，八城 美里，高瀬 裕，中野 有紀子，会話エージェントによる優位性推定に基づくグループ会話への介入，第 29 回人工知能学会全国大会，1I4-4, 2015