連想情報と時系列を考慮した文章の焦点推定による対話破綻検出

山﨑翔太　寺岡丈博
拓殖大学大学院 工学研究科 情報・デザイン工学専攻
20m309@st.takushoku-u.ac.jp　tteraoka@cs.takushoku-u.ac.jp

概要

現在の対話システムは，人間と円滑に対話できているとは言い難い．それは，人間と対話システム間における発話の意味理解が十分ではないからである．本研究では，対話中の話題を焦点と定義し，連想情報と時系列を加味した文章ベクトルで表した．この焦点とファインチューニングされた BERT の出力を特徴量として学習し得られた出力の結果，対話破綻検出において精度の向上を確認し，本研究の有効性を確認した．

1　背景と目的

近年，スマートデバイスの高性能化や移動通信技術の発展により，誰もが手軽に IT 技術を利用するようになった．しかし，IT 技術への慣熟度によって利用までのハードルの高さが違い，大きな格差が発生している．そこで，音声アシスタントやチャットボットによる操作案内等，IT 技術に不慣れでも IT 機器を操作できる技術が注目されている．

また，そのような流れの中で，非タスク指向型対話 (いわゆる雑談対話) の対話システム上への実装が模索されている．対話システムと人間との対話は，しばしば継続不可能な状況へ陥ることがある．そういった状況は，対話破綻 [1] と名付けられ，回避の手法が模索されている．

しかし，現状の対話システムはまだ発展途上であり，対話システムが話題の遷移を理解できず，対話破綻が引き起こされることも多い．そこで本研究では，対話が焦点としている概念の単語ネットワーク上での移動の推移を利用し対話破綻検出を行う．

2　先行研究

対話破綻の特徴に応じた回避手法の提案 [2] では，短期的な頻出語を話題と仮定し，それに関連する発話を行うことで対話破綻の回避を試みた．しかし，頻出語では話題推定の精度が低く，対話によっては元の発話よりも破綻感が高まってしまう等の結果が得られた．

松本・藤田らの研究 [3] では，Sugiyama らの研究 [4] の手法の再現性がない点に着目し，日本語 BERT モデルと他の特徴量を使った手法を再現性を持って実装し比較を行った．日本語 BERT モデルは対話破綻検出を破綻，軽微な破綻，破綻ではないの 3 種類に分類する文書分類タスクとして fine-tuning されている．その BERT モデルからの出力と，各発話の特徴量を結合した．低次元の特徴量を結合することで精度向上が期待できるということを明らかにした．

豊嶋の研究 [5] では，各発話文を Word2Vec によって発話ベクトルに変換し，素性を作成して SVM で対話破綻の検出を行っている．各発話の発話ベクトルの差分から得られたベクトル遷移の方向と大きさを利用したが，話題の遷移が見られた会話にのみ適応している等，課題点も見られる．

3　提案手法

3.1　概要

本研究では，出現した単語と，出現した単語を刺激語とした連想語の単語ベクトルを取得し，単語ネットワークを作成する．処理の流れを図 1 に示す．

作成したシステムは出現単語リストを持ち，それに操作を行うことで焦点を推定する．発話が行われると，発話された文章を分かち書きし，出現単語リストに追加する．また，追加された単語の連想される語を出現単語リストに追加する．この段階で，重み付けを行う．重みは連想されやすさと単語の特殊さを利用して算出する．その後，リスト内の単語の単語ベクトルを，Word2Vec を利用して取得する．

表 1　焦点の推移 類似度は焦点との類似度を指す

			Word2Vec 50 次元		Word2Vec 200 次元	
ターン数	発話者	発話	話題	類似度	話題	類似度
1	S	気持ちがえいですけどねぇ	気持	0.867	気持ち	0.737
2	U	潮風が気持ちいいし、海がきれい。	気持	0.804	気持	0.675
3	S	海は穏やかですね	海	0.848	海	0.821
4	U	昨日までの嵐が嘘みたいですね。	荒波	0.793	海	0.751
5	S	嵐は優しいですね	荒波	0.806	海	0.760
6	U	海岸を散歩しませんか？	海	0.913	海	0.883
7	S	海岸で貝殻を拾うのです	波打ち際	0.871	海	0.763

各単語の重みと取得したベクトルの加重平均を焦点ベクトルとする．また，システムは次の発話を受け取る前に，出現単語リストに付与された重みの減衰を行う．そして，重みが 0 未満となった単語を出現単語リストから削除し，発話待ち受け状態となる．単語ベクトルを取得する単語は名詞，形容詞，動詞に絞った．なお，「する」などの補助動詞は会話の焦点には関係ないと考え，除外した．ここで得られた加重平均を焦点，Word2Vec のモデル上で最も類似度の高い単語を話題として扱う．類似度は，コサイン類似度とした．

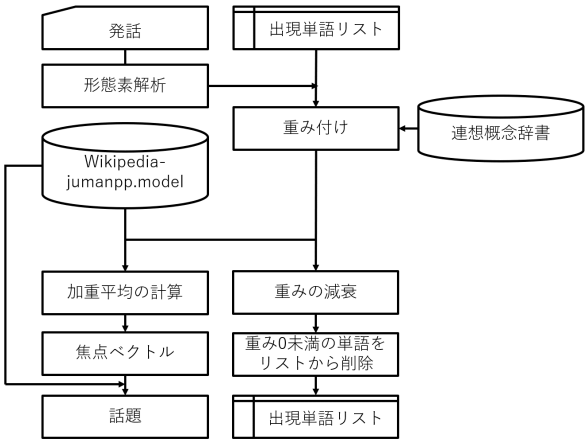

図 1　処理の流れ

3.2　焦点ベクトルの抽出

3.2.1　単語ベクトルの抽出

発話中の出現単語をリスト化し，リスト内に含まれる単語の単語ベクトルを wikipedia-jumanpp.model から抽出する．wikipedia-jumanpp.model は wikipedia を juman++で分かち書きし，出力された結果から学習させた Word2Vec のモデルである．

jumanpp は非常に精度が高い形態素解析器であるが，軽量と言われている形態素解析器である MeCab と比べて動作が遅い．そのため，Wikipedia 等の巨大なデータの分かち書きに使われる例は少ない．しかし，MeCab で分かち書きし学習させた Word2Vec モデルよりも高精度が期待できる．

3.2.2　興味値

表 2　変数の定義

説明	変数
出現によって加算される重み	N
連想された語の重み	RN ＝ N/連想距離
会話の進行で減衰する重み	SN ＝ N/M

会話の進行によって，会話中の概念への興味が変動すると推測し，各単語に重み付けを行った．今回の研究では，重みを興味値と定義した．興味値の算出は発話がされる度に行わる．各数値は，表 2 のように定義した．

人間が会話をする際，一度出現した単語でも会話が進むにつれて忘れられることから，何度も出てきた単語や連想されやすい単語，あまり一般的に利用されない単語が話題の焦点に近いと考えられる．そこで今回，対話中での出現回数，出現した単語との連想距離，日本語 Wikipedia 全文中の出現回数を利用して，以下のような方法で興味値を作成した．まず，単語の出現 (N) は，日本語 Wikipedia 上での出現回数を X とした場合に，N=1/logX で計算される．またこの時，出現回数 X が 10 未満であった場合は，出現回数 X=10 として扱う．これは，数ギガバイトの日本語 Wikipedia 上での出現回数は，一定回数以下であれば十分少ないと言えるからである．会話中で出現した単語については，N をそのまま興味値として扱う．連想距離は，連想概念辞書 [6][7] から抽出する．出現した単語を刺激語として連想された語は，N と出現した単語からの連想距離の逆数の積（N/連想距離）を興味値として加算する．これは，出現した単語が連想距離の原点にいると仮定して連想距離が遠いほど加算される重みを小さくするためである．N/連想距離は，RN とした．

また，興味値は加算が終了した後に減衰し，興味値が 0 未満になった単語は出現単語リストから削除される．興味値の減衰率は SN とした．SN は N/M であり，M ≦ N である．例えば，N=3,M=1 とした場合，1 度出現した単語は 3 発話分影響を及ぼすことになる．今回は，N=4,M=1 として実験を行なった．これは，後述する BERT の fine-tuning の際，破綻かどうかを判定されるターゲットになる発話を含めた直前 4 発話を利用するからである．

3.2.3 出力例

出力例を表 1 に示す．50 次元の Word2Vec では，海の話が進んでいるが，ターン数 4 の発話に含まれた「嵐」という単語から，焦点が「荒波」に寄ったことがわかる．次元数が低い場合，分散表現上の単語同士が近づくことで未出現の単語が話題とされやすい．

200 次元の Word2Vec では，ターン数 4 でも話題は海となっている．これは，次元数が増えることで，海と荒波の距離が広がり引き起こされたと考えられる．しかし，焦点との類似度は下がっており，話題が海から離れかけていることがわかる．

今回，Word2Vec の次元数は中間の 150 次元とした．これは，アンケートによって決定した．アンケートの参加人数は 5 名で，いずれも大学生となっている．

3.3 対話破綻検出

松本・藤田らの研究 [3] と同様に，対話破綻検出タスクに fine-tuning された BERT モデル [8] から得られた出力に，焦点ベクトルを結合することで行う．BERT モデルへの文章の入力は，BERT モデルを事前学習する時に利用したトークナイザでトークン化して行う必要がある．作成した焦点推定システム内では発話を jumanpp[9] で形態素解析しているため，同じ形態素解析器で形態素解析を行ってトークン化したモデルを利用する必要がある．そこで今回利用する BERT モデルは，黒橋・褚・村脇研究室から公開され日本語 BERT モデル [10] である．今回は，LARGE-WWM 版の BERT モデルを利用した．

3.3.1 BERT の fine-tuning

対話破綻検出を回帰問題として BERT の fine-tuning を行う．対話破綻検出チャレンジ [11] では，平均二乗誤差が評価指標として取り入れられている

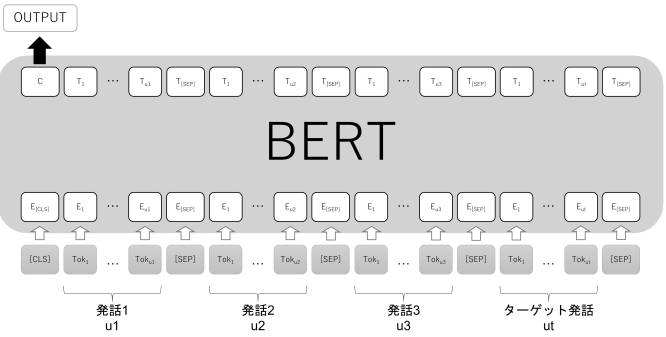

図 2 BERT モデルの概形

ため，fine-tuning 時の損失関数は，一般的にマルチクラス分類に利用されるクロスエントロピー誤差ではなく，平均二乗誤差を利用した．正解ラベルは，各発話での付与されたラベルの割合を利用する．例えば，20 人のアノテータがおり，そのうちの 10 人が破綻であると判断した場合は，破綻を表すラベルは"0.5"となる．対話破綻検出タスクに fine-tuning するために利用する発話は，ターゲット発話を含む直前 4 発話である．BERT モデルの概形は図 2 に示す．今回利用した BERT モデルは入力できる token 数が 128 に限られている．しかし，4 発話を用いるため，token 数が 128 を越える可能性がある．そこで，各発話を [CLS]token1 つと [SEP]token3 つが収まるように，31token になるように末尾を切り取った．また，最後の発話の token_type_id を 1 にすることで，ターゲット発話を明確化した．学習率は 9.5e-5 であり，epoch は 2000epoch 中で最も検証時の平均二乗誤差が低くなったモデルを利用した．学習の結果は表 3 である．

表 3 BERT fine-turning

	precision	recall	f1-score	support
破綻していない	0.76976	0.66191	0.71177	1192
破綻の可能性あり	0.37778	0.20238	0.26357	588
破綻している	0.46512	0.81081	0.59113	592
accuracy			0.58516	2372
Macro avg	0.53755	0.55837	0.52216	2372
MSELoss			0.04159	

3.3.2 誤差予測モデルの作成

誤差予測モデルは，中間層が 3 層，活性化関数が selu，第 3 層を dropout 層とした FFN を対話破綻検出タスクに fine-tuning された BERT モデルからの出力と，BERT モデルに入力した発話の特徴量を concat して学習データとしてトレーニングしたモデルである．図 3 今回，入力する特徴量を焦点ベクトルとすることで本システムの有効性を検証する．モデ

表 4　特徴量毎の結果　10 回平均

	焦点			BertVec			Word2Vec			
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	support
破綻していない	0.75406	0.69295	0.72160	0.76148	0.67718	0.71570	0.76703	0.66544	0.71111	1192
破綻の可能性あり	0.37011	0.20102	0.25532	0.37208	0.20425	0.25848	0.36894	0.19184	0.24619	588
破綻している	0.48443	0.77905	0.59602	0.47870	0.79426	0.59633	0.46752	0.81115	0.59264	592
accuracy			0.59250			0.58917			0.58440	2372
Macro avg	0.53620	0.55768	0.52431	0.53742	0.55856	0.52350	0.53450	0.55614	0.51665	2372
MSELoss			0.03991			0.04078			0.04174	

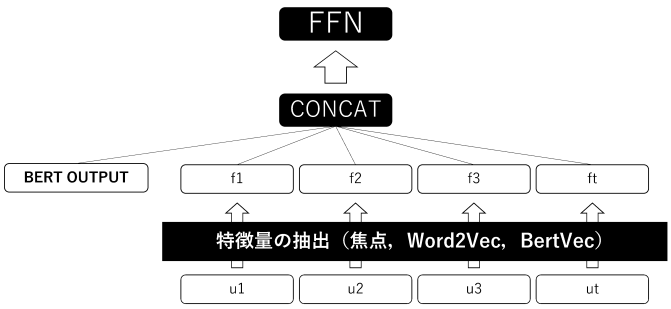

図 3　誤差予測モデルの概形

ルの概形は図 3 である．学習率は 6e-4，epoch 数は 20000epoch の中で最も検証時の平均二乗誤差が低くなった epoch のモデルを利用した．

3.3.3　実験設定

実験設定は DBDC4[12] に準拠した．学習データは DBDC4 の Development data として公開されているデータと，それ以前の対話破綻検出チャレンジ [11][13][14] の開発用データと検証用データを利用する．この中で，init1046 というデータに関してはアノテータが 2 人か 3 人と少なく，平均二乗誤差を損失関数として扱った場合，学習に悪影響を及ぼすため除外してある．検証データは，DBDC4 で検証データとして公開されたデータである，DBDC4_eval を利用した．また，誤差予測モデルの学習結果は学習ごとにかわるため，今回は 10 回学習を行い各数値の平均を結果として扱う．

3.3.4　結果

結果は表 4 のようになった．BertVec は BERT モデルの最終隠れ層の出力を各発話で平均した 768 次元のベクトルである．Word2Vec は，焦点と同じく 150 次元のモデルに単語埋め込みを行い，各発話で平均を計算した．

accuracy，MSELoss，”破綻している”の f 値が共に先行研究で使用された特徴量を上回った．また，ベースラインである BERT fine-tuning モデルを

4　考察

焦点を利用した対話破綻検出では，BERT fine-tuning モデルと比べて，"破綻している"の Recall が下がり，"破綻していない"の Recall が上がった．破綻した会話であると判定した数が増えたということは，焦点ベクトルが破綻していない会話の話題の推移を反映した特徴量であったからと考えられる．またこの結果から，破綻していない会話の焦点には，破綻した会話よりも特徴的な部分がある可能性が示された．しかし，対話破綻検出タスクでは，破綻する可能性が高い発話を確実に見つけられる方が，実際のシステムとして利用しやすい．破綻していない会話を破綻していると判断してしまった場合でも，より破綻していない発話を別で出力することになるだけであるためである．つまり本来は，"破綻している"会話を多く発見する必要がある．その場合は"破綻している"の Recall を向上させる必要があり，元の BERT の方が Recall が高く，目的を達成する可能性が高い．

焦点ベクトルを利用した対話破綻検出では，破綻類型によって結果が大きく変わる可能性が考えられる．例えば，ターゲット発話が非文であった場合の対話破綻は，焦点の推移による特徴量としては現れにくいと考えられる．また，発話の流れによる破綻でも，発話中に名詞，動詞，形容詞のいずれかが含まれない限りは，焦点推定の性質上直前の発話と焦点が大きく変わる可能性は低いため，特徴量としては現れにくい．

5　今後の課題

まず今後の課題として，焦点推定の精度向上が挙げられる．焦点推定を行う上でのパラメータがまだ定まっておらず，現状はタスク毎にヒューリスティックに決定している．焦点自体の定量的な評価は難しいため，タスク毎に定量的な評価を行いパラメータを設定するなどが考えられる．

謝辞

本研究は JSPS 科研費 JP18K12434, JP18K11514 の助成を受けたものです.

参考文献

[1] 東中竜一郎, 船越孝太郎. Project next nlp 対話タスクにおける雑談対話データの収集と対話破綻アノテーション. **SIG-SLUD**, Vol. 4, No. 02, pp. 45–50, 2014.

[2] 山崎翔太, 寺岡丈博. 対話破綻の特徴に応じた回避手法の提案. 第 82 回全国大会講演論文集, Vol. 2020, No. 1, pp. 421–422, 2020.

[3] 松本丈樹, 藤田桂英. Bert と文章の意味的特徴量を用いた誤差予測モデルによる対話破綻検出. Technical Report 8, 東京農工大学工学部, 東京農工大学大学院工学研究院, mar 2020.

[4] Hiroaki Sugiyama. Dialogue breakdown detection using bert with traditional dialogue. In **Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems**, Vol. 714, p. 419. Springer Nature, 2021.

[5] 豊嶋章宏. 応答の自然性と話題遷移を考慮した雑談対話システムにおける対話破綻の検出. 2018.

[6] 岡本潤, 石崎俊. 概念間距離の定式化と既存電子化辞書との比較. 自然言語処理, Vol. 8, No. 4, pp. 37–54, 2001.

[7] Takehiro Teraoka, Jun Okamoto, and Shun Ishizaki. An associative concept dictionary for verbs and its application to elliptical word estimation. In **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[9] 森田一, 黒橋禎夫. Rnn 言語モデルを用いた日本語形態素解析の実用化. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 13–14, mar 2016.

[10] 黒橋・褚・村脇研究室. Bert 日本語 pretrained モデル. **https://ja.wikipedia.org**.

[11] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将ほか. 対話破綻検出チャレンジ. **SIG-SLUD**, Vol. 5, No. 02, pp. 27–32, 2015.

[12] Ryuichiro Higashinaka, Luis FD'Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, João Sedoc. Overview of the dialogue breakdown detection challenge 4. In **Increasing Naturalness and Flexibility in Spoken Dialogue Interaction**, pp. 403–417. Springer, 2021.

[13] 東中竜一郎, 船越孝太郎, 稲葉通将, 荒瀬由紀, 角森唯子ほか. 対話破綻検出チャレンジ 2. **SIG-SLUD**, Vol. 5, No. 02, pp. 64–69, 2016.

[14] Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. Overview of dialogue breakdown detection challenge 3. 2017.