文脈および画像情報を加味した逐次的対話の解釈

大竹七勢 [1*] 松森匠哉 [1] 福地庸介 [1] 滝本佑介 [1] 今井倫太 [1]

[1] 慶應義塾大学 理工学部

Abstract: 人同士が対面で会話をする場合，会話中に登場した事柄や話題に触れながら話す文脈参照や，会話参加者の周囲に存在する物や出来事に触れながら話す状況参照が頻繁に起こり，時として双方が混在している．現状の対話システムにおける発話文処理では，ある発話に対して文脈と状況のどちらを参照対象とすべきか決定できず，文脈参照と状況参照の双方が混在する対話タスクを扱うことは難しい．本論文では，対話が逐次的に進む中で，文脈参照と状況参照それぞれに由来する解釈候補の中から発話のもっともらしい解釈を決定する手法 DICONV を提案する．

1 序論

人同士が対面で会話をする場合，会話中に登場した事柄や話題に触れながら話す文脈参照や，会話参加者の周囲に存在する物や出来事に触れながら話す状況参照が頻繁に起こり，双方が混在している．具体的な例として，人とロボットがスーパーで買い物をしている場面を挙げる．人が「魚と肉どっちが良い？」と聞いて，ロボットが「魚」と答えたとする．これはロボットが人の発話内容から質問の答えとなる語を参照しているので，文脈参照が起こっている．一方，目の前の商品棚に魚と肉が陳列されていて，人が「どっちが良い？」と聞いて，ロボットが「魚」と答えた場合には，ロボットが視覚情報から質問の答えとなる語を参照しているので，状況参照が起こっている．例で示した通り，実世界における対話では文脈参照と状況参照が起こり，両者は対話の中で混在する可能性がある．

文脈参照を扱う既存研究には，[1] や [2] などがあり，どちらも代名詞や指示語の参照対象を推定する照応解析タスクを扱っている．[1] はルールベースの手法であり，[2] は Deep Learning を利用したアプローチを取った手法である．状況参照を扱う既存研究には，[3] などがある． [3] は発話上から抽出されたオブジェクトと画像中から検出されたオブジェクトの類似度をスコア化し，画像中の参照対象を推定した．

文脈参照と状況参照を扱う既存研究に共通する問題点は，対話における文脈参照と状況参照のどちらか一方しか扱っていないことである．実世界の対話では，文脈参照が発生するか，状況参照が発生するかは発話時の場所や状況などに依存し，予め決定することはできない．現状の対話システム技術では，ある発話に対して文脈と状況のどちらを参照対象とすべきか決定できず，文脈参照と状況参照の双方が混在する対話タスクを扱うことは難しい．そのため，上に示した既存研究のように，通常は，チャットボットに代表される文脈参照が多い場面と，ロボットへの指示タスクに代表される状況参照が多い場面のどちらかを予め仮定し，対話システムを設計する．よって，文脈参照と状況参照を統合的に扱ったモデルは存在しない．人とロボットとの買い物など，人と対話システムが話し合いながら協力してタスクをこなすためには，対話において文脈参照か状況参照か曖昧な状態の中で，現在の発話がどちらに当たるのか逐次的に推定した上で発話を決める必要がある．

本論文では，対話が逐次的に進む中で，文脈参照と状況参照それぞれに由来する解釈候補の中から発話のもっともらしい解釈を決定する手法 DICONV(Dynamic and Incremental Interpretation of Contextual and Visual References in Conversational Dialogues) を提案する．DICONV は，逐次的に対話が進む中で，文脈参照を考慮した発話解釈候補の確率的な探索と，状況参照を考慮した発話解釈候補の確率的な探索を同時に行い，それぞれの解釈結果を比較して，どちらか一方の参照を選択する．このようにして，文脈参照と状況参照の双方を同時的に扱うことを可能にする．

本論文の構成は次の通りである．次章では，本論文の提案手法 DICONV について詳細に述べる．3 章では，本論文で行った実験の概要や実験の結果，その結果に対する考察について述べる．4 章では本論文全体のまとめを述べる．

2 DICONV

逐次的な対話において文脈情報と視覚情報の双方を考慮しつつ文脈と単語の解釈を動的に推定する手法として，Dynamic and Incremental Interpretation of Contextual and Visual References in Conversational Dia-

*連絡先：慶應義塾大学
E-mail: otake@ailab.ics.keio.ac.jp

logues を提案する．DICONV では，逐次的な対話に出現する指示語の参照対象が文脈参照に由来するものか，状況参照に由来するものかを決定し，指示語の参照対象を推定することが可能である．

逐次的な対話が進む中で，指示語の複数の発話解釈候補を確率的に探索する手法として，既存研究である SCAIN[4] を用いる．

3 実験と考察

3.1 実験設定

逐次的な対話における指示語推定タスクで，提案手法 DICONV を評価した．指示語の参照先が文脈参照であるか状況参照であるかを決定し，指示語の解釈候補を推定することが可能であるのかを検討することを目的として，以下の 2 つの発話例 (表 1, 2) に関して実験を行った．各発話例に関して，どちらも「スーパーでの買い物中の会話」という状況を想定したものである．これらの指示語について，文脈参照であるか状況参照であるかが適切に決定できれば，文脈参照と状況参照が混在した対話を扱うことが可能であると判断する．

表 1: 発話例 1（指示語の参照先：lemon，文脈参照）

発話者	発話内容
A	“I have a cold. Do you know any fruit that is good for colds? ”
B	“How about kiwi and pineapple? ”
A	“Well, anything else? ”
B	“How about lemon? ”
A	“Oh, it is certainly effective. ” (∗ 文脈参照・状況参照ともに解釈候補を取得)
B(入力 1-1)	“Yeah, It is rich in vitamin C and citric acid, so I think it is effective.”
A	“OK, I must buy it.”
B(入力 1-2)	“Yeah, but, It is sour and hard to eat, so I recommend eating it with other foods.”

表 2: 発話例 2（指示語の参照先：banana，状況参照）

発話者	発話内容
A	“I have a cold. Do you know any fruit that is good for colds? ”
B	“How about lemon, kiwi and pineapple? ”
A	“Good. How do you think of that fruit? ” (∗ 文脈参照・状況参照ともに解釈候補を取得)
B(入力 2-1)	“It contains many types of sugars and changes quickly to energy. So, maybe it is effective. ”
A	“Any other information about that fruit? ”
B(入力 2-2)	“It is perishable, so you need to eat them early. ”

ここでは，各発話例の最初に出現した指示語を解釈対象とし，以降に出現する指示語は解釈対象の指示語と同じものを参照しているとして候補を推定した．指示語の解釈候補に関して，各発話例の指示語が出現したタイミングで，それ以前の発話から適切な単語を抽出し，これを文脈参照による解釈候補とした．また，同時に画像も取得し，YOLO[5] によって画像から抽出されたラベルを状況参照による解釈候補とした．具体的な解釈候補はどちらの発話例にも共通で，文脈参照による解釈候補が lemon, kiwi, pineapple であり，状況参照による解釈候補が banana, apple, orange である．これらの解釈候補を DICONV の解釈候補として設定し，最初の指示語が含まれる発話以降の発話を逐次的に入力し，指示語の推定を行った．

また，表 1, 2 中にも示したが，各発話例の正解は，発話例 1 の指示語の参照先が文脈参照による解釈候補である lemon, 発話例 2 の指示語の参照先が状況参照による解釈候補である banana を想定している．

3.2 実験結果

DICONV に逐次的に会話内容を入力した結果，各発話例においてターゲットと各解釈候補とのコサイン類似度は次のように変化した．表 3 中の入力 1-1, 1-2, 表 4 中の入力 2-1, 2-2 は表 1, 2 にそれぞれ示した文章に対応している．表 3, 4 中のコサイン類似度の値は，初期位置と各入力の後に算出された最良のパーティクルに関してのものである．ここで，初期位置に関して，何も入力がない状態では，パーティクル間の重みに差は

ないので，どの解釈候補も確信度は等しくなる．コサイン類似度を確信度と同等と捉え，各解釈候補が参照対象かそうでないかが分からない状態を示す 0.5 を初期位置の値とした．

表 3: 発話例 1 における文脈参照・状況参照の各解釈候補とターゲットとのコサイン類似度

解釈候補 / 入力	初期位置	入力 1-1	入力 1-2
文脈参照			
lemon	0.5000	0.9962	0.9989
kiwi	0.5000	0.3458	0.3432
pineapple	0.5000	0.7201	0.6979
状況参照			
banana	0.5000	0.5632	0.9236
apple	0.5000	0.5383	0.5710
orange	0.5000	0.9820	0.7643

表 4: 発話例 2 における文脈参照・状況参照の各解釈候補とターゲットとのコサイン類似度

解釈候補 / 入力	初期位置	入力 2-1	入力 2-2
文脈参照			
lemon	0.5000	0.9960	0.9699
kiwi	0.5000	0.3504	0.3776
pineapple	0.5000	0.7206	0.8383
状況参照			
banana	0.5000	0.9784	0.9934
apple	0.5000	0.5245	0.5167
orange	0.5000	0.5893	0.5276

推定結果は，発話例 1 では表 3 中の入力 1-2 の列で最もコサイン類似度が大きいものである．発話例 2 では表 4 中の入力 2-2 の列で最もコサイン類似度が大きいものである．よって，発話例 1 は文脈参照の解釈候補 lemon, 発話例 2 は状況参照の解釈候補 banana である．どちらも想定した指示語の参照対象を得られている．

3.3 考察

発話例 1 に関して，表 3 から，初期位置を除いてコサイン類似度が最も大きい解釈候補をたどると，文脈参照は lemon → lemon, 状況参照は orange → banana となっている．発話例 1 で想定した指示語の参照対象は lemon なので，発話を適切に解釈し，安定した推定ができているといえる．同様に，発話例 2 に関して，表 4 から，文脈参照は lemon → lemon, 状況参照は banana → banana となっている．どちらも安定した推定となっている．表 4 中の文脈参照と状況参照それぞれの最も大きいコサイン類似度を持つ候補 lemon, banana のみに着目し比較する．表 4 から，入力 2-1 から入力 2-2 の間に banana はコサイン類似度が大きくなっているが，lemon は小さくなっている．発話例 2 で想定した指示語の参照対象は banana なので，発話を適切に解釈できているといえる．

4 結論

既存研究で未解決な課題は，逐次的な対話において文脈参照と状況参照が混在した場面を扱えないことであった．

本論文では，既存研究である SCAIN を用いて，文脈参照と状況参照を並列的に処理し，最終的な推定において双方の推定結果を統合することで未解決な課題の扱いを可能とした．

実験では，文脈参照と状況参照の双方を含む発話例に関して，指示語の参照対象の推定を行い，文脈参照と状況参照の混在したタスクを扱えることを確認した．

参考文献

[1] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pp. 28–34. Association for Computational Linguistics, 2011.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3774–3781. IEEE, 2018.

[4] Yusuke Takimoto and Michita Imai. Slam-inspired simultaneous contextualization and interpreting for conversation sentences. 2019.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.