視線情報を利用した協調作業対話における参照解析

安原 正晃　　飯田 龍　　徳永 健伸

東京工業大学 大学院情報理工学研究科

1　はじめに

人と協調して作業するシステムが幅広い利用者に受け入れられるためには，複雑な操作手順を覚える必要がなく，人間が日常で用いる自然言語によってコミュニケーションできることが必要である．とりわけ，協調作業におけるコミュニケーションでは，ある物を他の物と混同せずに指し示す参照表現を適切に使う能力が求められる．つまり，システムは人間と自然言語でインタラクションする過程で，参照表現を使用したり (参照表現生成)，人間の使用する参照表現が何を指してるのかを理解する (参照解析) 必要がある．しかし，参照表現を利用する際に重要となる情報は状況に依存するため，システムにこのような能力を持たせるためには様々な状況下で分析を行なう必要がある．我々は，2 名で協調してパズルを解くという協調作業を取り上げ，その中で用いられる参照表現について生成・解析の両面から研究を行なっている [3, 11].

参照解析については，先行研究 [3] において，対話中の文脈から得られる談話履歴の情報に加えて，直示，操作履歴といった非言語的な情報を用いることで，解析の精度が向上することを明らかにした．特に，代名詞と代名詞以外の参照表現を分けて解析を行なうことで，代名詞において，最大 0.238 (0.648 → 0.886) の精度の向上が見られた．しかし，代名詞以外の参照表現では，最大でも 0.014 (0.680 → 0.694) の精度向上しか得られなかった．本稿では，参照表現解析の性能をさらに改善するために，対話中の話者の視線情報を利用した結果について報告する．

視線情報は，人間の興味・関心・意図を強く反映している情報であり，人間への負担をかけることなく利用可能な情報である [5, 8, 6]．そこで本稿では，この視線情報を用いて，協調作業対話において人間の使用する参照表現を解析する手法を提案する．

2　協調作業対話コーパス

本研究で使用した参照表現コーパスは，被験者に 2 名 1 組で図形パズルを協力して解く課題を与え，収集したものである [12]．この課題は，提示した「目標図形」 (図 1 左側) と同じ形を「作業領域」 (図 1 右側) にあるパズル・ピースを用いて作成する課題である．1 組の被験者は，それぞれ「指示者」と「作業者」という役割を割り当てる．指示者は，目標図形を構成するパズル・ピースの配置を考え，ピースの操作を作業者に指示する．作業者には目標図形が見えないので，指示者の指示にしたがって，パズル・ピースをマウスで操作する．なお，指示者はマウス操作ができないので，自分で直接ピースを操作することはできない．指示者と作業者は，音声とシミュレータの作業領域のリアルタイムの映像によってのみインタラクションが許されるが，対話の内容に関しての制限はない．

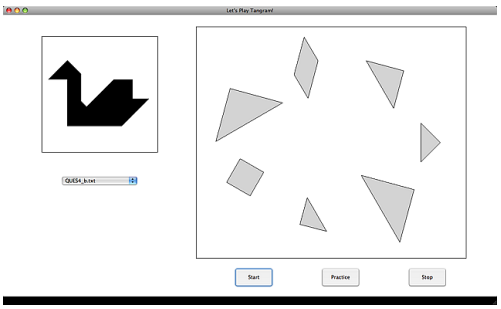

図 1: 図形パズル・シミュレータ

表 1: ELAN で管理するアノテーション層

層の名前	意味
OP-UT	作業者の発話
SV-UT	指示者の発話
OP-REX	作業者の使った参照表現
OP-Ref	その指示対象
OP-Attr	その構成属性
SV-REX	指示者の使った参照表現
SV-Ref	その指示対象
SV-Attr	その構成属性
Action	ピースに対する操作
Target	その操作の対象ピース
Mouse	マウスカーソルの位置
OP-GZE-P	作業者の注視点
OP-GZE-N	作業者の注視点の最近傍ピース
SV-GZE-P	指示者の注視点
SV-GZE-N	指示者の注視点の最近傍ピース

※ 先頭の字下げはアノテーション層の親子関係を表わす．

作業中の被験者の発話音声 (指示者：SV-UT，作業者：OP-UT)，ピースおよびマウス操作 (Action, Mouse: 1/65 秒間隔)，視線位置 (*-GZE-*: 1/60 秒間隔) を時間同期して記録し，アノテーション・ツール ELAN [1] によって対話中の参照表現をアノテーションした．アノテーション情報の一覧を表 1 に示す．なお，視線位置の計測は Tobii 社の T60 を用いた．

本稿で利用した協調作業対話コーパス中の 2 つのコーパス (T2009-11 と N2009-11) の概要を述べる．T2009-11 は 7 組 27 対話 (総対話時間 4:22:20)，N2009-11 は 2 組 8 対話 (総対話時間 1:47:45) を収録している．対話中の参照表現の総数は，T2009-11 で指示者 1,192 表現，作業者 270 表現，N2009-11 で指示者 497 表現，作業者 168 表現である．このコーパスには作業領域中のパズル・ピースを指示対象とする各参照表現が，指示対象の ID (*-Ref) と表現中に含まれる属性 (*-Attr) と共にアノテーションされている．本稿では，指示者が使用した参照表現のうち，指示対象として 1 つのパズル・ピースを指すもののみを対象した．その数は，T2009-11 で 1,093 表現，N2009-11 で 390 表現である．

表 2: 素性の一覧

D1	最後に言及されたピースか
D2	最後に言及されてからの経過時間が 10 秒未満のピースか
D3	最後に言及されてからの経過時間が 10 秒以上 20 秒未満のピースか
D4	最後に言及されてからの経過時間が，20 秒以上のピースか
D5	以前に一度も言及されていないピースか
D6	参照表現の持つ属性 (形・大きさ) が，ピースの属性と矛盾しないか
D7	最後にそのピースを指す表現がヲ格として用いられているか
D8	最後にそのピースを指す表現が二格として用いられているか
D9	参照表現が代名詞の時，その直前にピースが代名詞以外の表現で参照されているか
D10	参照表現が代名詞以外の表現の時，その直前にピースが代名詞で参照されているか
M1	発話開始時刻にオンマウスされているピースか
M2	発話開始時刻にオンマウスしていないとき，直前にオンマウスしていたピースか
M3	最後にオンマウスされてからの経過時間が 10 秒未満のピースか
M4	最後にオンマウスされてからの経過時間が 10 秒以上 20 秒未満のピースか
M5	最後にオンマウスされてからの経過時間が 20 秒以上のピースか
M6	以前に一度もオンマウスされていないピースか
A1	発話開始時刻に操作されているピースか
A2	発話開始時刻に操作されていない場合，直前で操作されているピースか
A3	最後に操作されてからの経過時間が 10 秒未満のピースか
A4	最後に操作されてからの経過時間が 10 秒以上 20 秒未満のピースか
A5	最後に操作されてからの経過時間が 20 秒以上のピースか
A6	以前に一度も操作されていないピースか
G1	発話前一定期間の注目頻度
G2	発話前一定期間の注目時間 (抽出区間で正規化)
G3	発話前一定期間の注目時間 (総注視時間で正規化)
G4	発話中の注目頻度
G5	発話中の注目時間 (抽出区間で正規化)
G6	発話中の注目時間 (総注視時間で正規化)

3 参照解析手法

参照表現の同定にはランキング・モデルに基づく手法を用いる．ランキング・モデルとは，指示対象候補すべての中でどれがもっとも指示対象らしいかを，ランカーを用いてランク付けし，その 1 位を指示対象と判定する多値分類をおこなうモデルである．具体的には，以下のように参照解析をおこなう．

(1) 各参照表現について，その表現の発話開始時までの各ピースの状況を，表 2 の素性を用いて表現し，ピースごとに特徴ベクトルを作成する．

(2) Ranking SVM [4][1]を用いて，訓練データからランカーを作成する．2 値分類を行なう通常の SVM に対し，Ranking SVM は指定したグループ内でのランク付けを行なうことができる．ランカーの学習では，人手でタグ付けをした指示対象を 1 位，その他のピースを 2 位として，学習をおこなう．

(3) (2) 作成したランカーを利用し，テストデータの各ピースの特徴ベクトルをランク付けする．

(4) (3) で 1 位となった特徴ベクトルに対応するピースを，その表現の指示対象とする．

Iida ら [3] が用いた素性 (談話履歴情報，オンマウス情報，操作履歴情報) に加え，視線情報の素性を用いる (表 2)．談話履歴情報 (D) は，従来の照応解析手法と同様に，談話の先行文脈から得られる情報を素性として用いる．オンマウス情報 (M) は，ELAN の Mouse 層の情報を使い，作業者の操作するマウスがピース上に乗っている状態をオンマウス状態として，これを素性として用いる．オンマウス状態は，人間同士の実環境における直示 (指差しなど) の情報に近いものと考えることができる．操作履歴情報 (A) は，ELAN の Action 層の情報を参照し，作業者が何らかのピースを操作しているという情報を利用する．本環境では，作業者はピースの移動，ピースの回転，ピースの反転の 3 つの操作を選択しておこなうことができるが，ここではこの 3 つの操作を区別せず，操作を行なっているか否かという粒度でこの情報を扱う．

本稿で導入する視線情報 (G) は，視線が人の注目情報を表す [5, 9] という知見に基づいている．物体を指示する場合には，その物体へ注目しており，視線もその物体へ向いていると考えられる．コーパスでは，注視を「許容誤差の範囲に収まる連続した点の集合」と定義し，*-GZE-P 層に注視の中心座標 (注視点) が記録されている．許容誤差は，T60 のカタログ仕様の測定誤差 (0.5 度)，画面から被験者までの平均的な距離 (50cm)，および，画面解像度 $(1,280 \quad 1024)$ から計算し，16 ピクセルとした．また，Richardson らの実験 [9] にならい，視線がこの許容誤差内の領域に 100m 秒以上留まる時に注視であるとみなした．ある時区間 (抽出区間) で発生した各ピースの注目に基づいて視線素性 (G1–G6) を以下のように計算する．ここでピースへの注目はピースが注視点から許容偏差内に含まれる事象をいう．人間の有効視野 (parafoveal visual field) の範囲は約 20 度であると言われていることから [7]，注目のモデルとして，注目しているかどうかを $[0, 1]$ のスコアの値で返す以下の 4 つのモデルを考える．

[1] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

- 最近傍注目モデル (Nst)：有効視野中にある注視位置から最近傍にあるピースは 1，それ以外は 0 を返す．
- 一定範囲注目モデル (Fix)：有効視野中にあるピースは 1，それ以外は 0 を返す．
- 単調減少注目モデル (Dec)：有効視野中にあるピースは注視位置からの距離に応じて線型に減少するスコアを返し，有効視野外にあるピースは 0 を返す．
- 段階範囲注目モデル (Vec)：有効視野を半径 1 度刻みで 11 段階に分け，それぞれその領域内に含まれるピースには 1，含まれないピースには 0 を与えて作る 11 次元のベクトルを返す．

これらの注目モデルを用いて注目頻度 (G1，G4) および 注目時間 (G2, G3, G5, G6) を考える．注目頻度は抽出区間中に発生した総注視回数で正規化する．注目時間は，抽出区間長で正規化する場合 (G2, G5) と，抽出区間中の注視の合計時間で正規化する場合 (G3, G6) の 2 つを考える．つまり，ある抽出区間について，

- $\mathrm{G1(G4)} = \sum \frac{\mathrm{Att}}{\text{抽出区間中に発生した総注視回数}}$
- $\mathrm{G2(G5)} = \sum \frac{\mathrm{Att}\times\text{注視時間}}{\text{抽出区間長}}$
- $\mathrm{G3(G6)} = \sum \frac{\mathrm{Att}\times\text{注視時間}}{\text{抽出区間中に発生した総注視時間}}$

となる．ここで Att は注目モデルが返す値である．また，抽出区間としては，参照表現発話前の一定区間 (G1–G3)，と発話中 (G4–G6) を考える．なお，段階範囲注目モデルは各視野角について二値の値を返すので，素性は見掛け上 11 倍となる．

4 評価実験

4.1 実験設定

2 節で説明したコーパスから，N2009-11 を視線情報の抽出範囲決定のための開発セット，T2009-11 を評価セットとして使用する．それぞれに出現する指示者が発話する指示対象がピース 1 つの参照表現 390 表現 (N2009-11) と 1,093 表現 (T2009-11) を使用した．ただし，T2009-11 を使用した評価実験の際は．代名詞に着目し，代名詞を含む表現 409 表現と代名詞を含まない表現 684 表現を分け，個別に評価実験をおこなう．これは，代名詞が他の参照表現と比較して直示や先行詞との時間的な近さの影響を受けやすい点を考慮したためである [10]．特徴の異なる代名詞とそれ以外を区別し，それぞれの特徴に合ったランカーを作成することで，精度の向上を図った [2]．以上の理由から，代名詞 (Pro) と代名詞以外の参照表現 (non-Pro) は個別に学習・評価する．

表 3: 各素性の組合せの精度と正解数

	BL	+G(Nst)	+G(Fix)	+G(Dec)	+G(Vec)
Pro	0.800	0.800	0.790	0.800	0.804
(409)	327	327	323	327	329
non-Pro	0.675	0.749	0.677	0.740	0.746
(684)	462	512	463	506	510

4.2 視線情報抽出区間の決定

抽出区間を決定するために，N2009-11 コーパスを用いて，注目頻度 (G1) と注目時間 (G2, G3) が指示対象のピースについて最大のスコアとなる時区間を求めた．なお，注目についは最近傍注目モデルを用いた．発話前 100m 秒から発話前 3,000m 秒まで 100m 秒区切りで抽出範囲を変化させて比較を行なった結果，発話前 1,300m 秒を抽出区間とした場合が最も指示対象ピースの注目頻度，注目時間が高くなる．この結果から，評価実験の素性 G1–G3 については発話前 1,300m 秒を抽出区間とした．

4.3 実験結果

実験結果を表 3 に示す．列は解析に使用した素性の組合せ，行は解析対象の参照表現 (Pro：代名詞，non-Pro：代名詞以外) に対応する．また，各セルの上段は精度，下段は正解数である．ベースライン (BL) としては，Iida ら [3] が用いた素性の組合せ (D，DM，DA，DMA) のうち，評価セットに適用して精度が最大になる組合せ用いた．提案手法では，DMA の素性組合せに 4 つの注目モデルを用いた視線情報をそれぞれ加えた素性の組合せ (G(Nst)，+G(Fix)，+G(Dec)，+G(Vec)) を考える．

代名詞の解析では，ベースラインの DM の素性組合せに対して，+G(Vec) の素性組合せで，0.004 とわずかながら精度が向上した．代名詞以外の解析では，ベースラインの DMA の素性組合せに対して，+G(Nst) の素性組合せで，0.074 の精度が向上した．

表 4: 性能改善の McNemar 検定結果

Pro	解析対象	BL(DM)	DMAG(Vec)	p
	全表現	0.800	0.804	0.724
	視線有	0.807	0.818	0.465
non-Pro	解析対象	BL(DMA)	DMAG(Nst)	p
	全表現	0.675	0.749	4.4e-07
	視線有	0.673	0.750	3.3e-07

ベースラインと提案手法の有意差を検証するために McNemar 検定を行なった．ベースラインと比較する提案手法には，最も精度が高かった注目モデルを用いる．表 4 に McNemar 検定により求めた有意確率を精度と共に示す．解析対象とした参照表現の中には，発話近傍に注視が形成されず，視線情報が使用できないものも含まれる．そこで，発話近傍で注視が発生し，視線情報が利用可能な表現のみを「視線有」行に併記した．視線情報が利用可能な表現は代名詞で 357 表現（87%），

表 5: 各素性カテゴリ単独の精度と正解数

	D	M	A	G (Best)
Pro	0.560	0.790	0.653	0.531
(409)	229	323	267	217
non-Pro	0.658	0.222	0.151	0.433
(684)	450	152	103	296

代名詞以外で 645 表現（94%）である．代名詞の解析では，有意確率は全表現，視線情報が利用可能な表現ともに大きな確率を取り，有意水準を 1%未満とすると，有意な差はみられない．一方，代名詞以外の参照表現の解析では，有意確率は全表現，視線情報が利用可能な表現ともに 1%を下回り，有意水準 1%未満で有意な差があった．

表 5 に示す素性カテゴリごとの単独の解析精度を見ると，代名詞の解析では，Iida らの手法 [3] は，オンマウス情報 M が単独で解析精度 0.790 とかなり高い精度を示し，ベースラインや提案手法の性能の大部分を担っている．オンマウス情報 M に比べると，視線情報 G(Vec) 単独での精度は低いため，DMA の素性カテゴリに，さらに視線情報を加えても，解析精度はそれほど向上しなかったと考えられる．一方，代名詞以外の参照表現の解析では，先行研究で精度の高かった素性カテゴリ (M, A) に比べて，視線情報 G(Nst) 単独での解析精度が高い．そのため，視線情報を加えた本提案手法がベースラインに比べて，有意水準 1%未満の高い精度の向上がみられたと考えられる．

5 おわりに

本稿では，協調作業対話における参照表現の指示対象の解析に視線情報を導入する手法を提案した．具体的には，先行研究で用いた談話履歴，直示，操作履歴などの素性に加え，視線情報を素性として追加し，機械学習の手法を利用した．提案手法を用いて評価実験を行なった結果，先行研究に比べ，特に代名詞以外の参照表現について精度が最大で 0.074 向上 (0.675 → 0.749) することを確認した．

今後の課題として，まず，別領域での評価が考えられる．本稿では，協調作業対話コーパス中のタングラム・パズルを題材としたコーパス (T2009-11 と N2009-11) を用いた．本手法が別の領域でも有効に働くことを確認するために，他のコーパスでも検証する必要がある．

さらに，解析モデルを個人適応することも考えられる．本稿では，複数の被験者の視線情報を利用して，平均的な解析モデルを作成した．しかし，視線の動きや有効視野には個人差があることが知られている．また，視線計測における視線計測率などにも個人差が生じるため，個人間で視線情報の有効性は異なる．そのため，複数被験者の視線を用いて作成した平均的モデルを個人適応すれば, さらに精度が向上する可能性がある．

謝辞

本研究は研究費補助金基盤研究 (B) 2130049 の助成によるものである．

参考文献

[1] Eric Auer, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, S. Masnieri, Daniel Schneider, and Sebastian Tschöpel. ELAN as flexible annotation framework for sound and image processing detectors. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp. 890–893, 2010.

[2] Pascal Denis and Jason Baldridge. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 660–669, 2008.

[3] Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1259–1267, 2010.

[4] Thorsten Joachims. Optimizing serch engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

[5] Marcel Adam Just and Patricia A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, Vol. 8, pp. 441–480, 1976.

[6] Mary C. Potter. Representational buffers: The eye-mind hypothesis in picture perception, reading, and visual search. In Keith Rayner, editor, *Eye movements in reading: Perceptual and language processes*, pp. 423–437. Academic Press, 1983.

[7] Keith Rayner. Parafoveal idenfication during a fixation in reading. *Acta Psychologica*, Vol. 39, No. 4, pp. 271–281, 1975.

[8] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, Vol. 124, No. 3, pp. 372–422, 1998.

[9] Daniel C. Richardson, Rick Dale, and Michael J. Spivey. Eye movements in language and cognition: A brief introduction. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael J. Spivey, editors, *Methods in Cognitive Linguistics*, pp. 323–344. John Benjamins., 2007.

[10] Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*, 2009.

[11] Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Terai Asuka, and Kuriyama Naoko. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources & Evaluation*, 2010.

[12] 安原正晃, 石川真也, 飯田龍, 徳永健伸. 視線情報を含むマルチモーダル協調作業対話コーパスの構築と利用. 情報処理学会自然言語処理研究会, 第 NL-199-20 巻, 2010.