共参照解析のための 英語対話形式テキストの自動書き換え

藤川 哲志　　但馬 康宏　　菊井 玄一郎
岡山県立大学大学院　情報系工学研究科
{cd25044k，tajima，kikui}@cse.oka-pu.ac.jp

1. はじめに

共参照解析は，テキストの意味を理解する上で必要不可欠な処理である．共参照解析は，その結果自体がテキスト理解の出力を構成するだけでなく，述語項構造解析を始め感情解析，発話意図推定など他の処理を正しく行うためにも必要な基盤技術である．このため，共参照解析は，多くの研究が行われてきた([1][2]など)．しかしながら，現状においても広範なテキストについて満足できる性能にはなっていない．実際，本研究で対象としている対話形式のテキスト(話者名のあとに台詞が続く形式)については，2011年に行われたCoNLLの共参照Shard Task[1]で最も良い結果を出したStanford CoreNLP[2]（以後CoreNLPと表記)でも基本的なところで誤ってしまう．

この問題に対して，本研究では，対話形式のテキストを「普通の」テキスト(ここでは「叙述形式」と呼ぶ)に自動変換することにより，CoreNLPによる共参照解析の精度向上を目指す．対話形式のテキストとして，英語のセンター試験で出題される対話の穴埋め問題（以後対話文問題と表記）の空所に正解の選択肢を入れたもの(以後対話テキストと表記)を用いる．なお，本研究は，英語センター試験の自動解答[3]への応用を想定しているが，通常の対話形式のテキスト全般に対して適用可能なものである．

2. 対話テキスト

本研究で対象とする対話テキストとは，図1と図2に示すように，一つの発話(ターン)を，発話者の名前(以後，「発話者記述」と表記)をX，発話内容(「発話」と表記)

Mr. Abbott: Hello, is Mr. Pratt in?
Secretary: I'm sorry, but he is out of town on a business trip.
Mr. Abbott: I see. Then, may I speak to Ms. Lee ?
Secretary: Let me check.

図 1 対話テキストの例
(2007年センター試験追試験第2問B問1)

A: Your French is very good, Jane! (1)
B: Thank you. It's kind of you to say so. (2)
A: Did you study in France? (3)

図 2 発話者が変数(A，B)である例
(2001年センター試験本試験第2問B問3)

をSとすると「X: S」の形式で記述した，芝居の台本のような形式のテキストである．発話者記述は，Mr. Abbottのような「固有名」，Secretaryのような「役割名」，AやBといった「変数名」のいずれかである．図2の下線部のように発話中で次の発話者の名前が“Jane”と明示されているにも関わらず発話者記述がそれと異なる“B”となっている場合がある．この場合，何らかの方法でB=Janeと解釈する必要がある．

3. 既存手法の問題点

3.1. CoreNLPの共参照解析

CoreNLPは，品詞タグ付けや構文解析，固有表現解析，共参照解析などの言語処理ツールをまとめたパッケージである．CoreNLPの共参照解析器は，テキスト中のmentionを抽出したあと，談話情報や構文，固有名詞などの情報に基づいて，それらの間の共参照関係の有無を判定する，ルールベースのシステムである．

3.2. 基本的な問題点

CoreNLPに対して図1のような対話テキストをそのまま入力した場合の共参照解析の精度を表3(5.1.節)の「書き換えなし」の行に示す．精度が低い最大の理由は，mention抽出が不適切なことにある．具体的には，「:」をしばしば同格的に捉えて，発話内容を発話者記述の説明文と解釈し，発話者記述を含む発話のテキスト全体を1つのmentionとして出力する場合がある．

4. 提案手法

4.1. アプローチ

そこで，本研究では発話ごとに，直接話法の文に変換して，対話形式でない形にしてからCoreNLPを適用する手

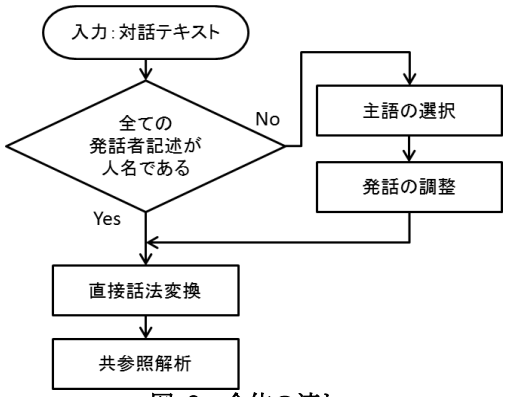

図 3　全体の流れ

法を提案する. 直接話法の文にする基本的な方法は,「X: S」という形式を「X said,"S"」の形式に書き換えることである．しかし，発話者記述 X が Secretary のような「役割名」や A，B のような「変数名」の場合，X の部分をそのまま直接話法の文の主語にすると CoreNLP が X を mention として認識しないことがある. そこで, これらの場合は，主語を固有名詞(人名)等に変換する．その際，図 2 のような現象を扱うために発話中の人名を抽出・利用する．

4.2.　全体の流れ

図 3 に提案手法の全体の流れを示す. 対話テキストを入力とし, すべての発話者記述が人名であるか名前辞書を用いて確認する．すべて人名であるなら，ターン毎に直接話法に変換し，共参照解析を行う．一つでも人名でないものがあれば，役割名か変数名であるので，直接話法の文の主語を選択したあと，直接話法の文に変換を行い，共参照解析を行う．以下各処理と辞書の構築について説明する．

4.3.　直接話法変換

提案手法の基盤となる処理である．松尾[4]によると，直接話法とは,「<発言者>の発話を他人のことばとして区別し，<報告者>が客観的に報告する話法」である．つまり, ある発話者が発言したことばを第三者に報告するための話法で，発話の時制などを考慮する必要がない．直接話法では，伝達部(X said,)は発話のどの位置でも置くことが可能であるので，本研究では先頭に置く．また伝達動詞は，tell や speak などがあるが，say の過去形 said で統一する．例を以下に示す.

（変換前）Mr. Tani: What time did you come?
（変換後）Mr. Tani said, ``What time did you come?"
(2013 年センター試験本試験第 2 問 B 問 1)

直接話法の文は，CoreNLP に談話構造のルールとして存在しているため，精度の高い処理が期待できる．

表 1　パターンリスト

パターン名		パターン	PERSONの指示対象
電話		this be PERSON	発話者自身
自己紹介		my name be PERSON my name 's PERSON	
呼びかけ	(Final)	…, PERSON 文末記号	次の発話者
	(Initial)	PERSON , …	
	(Stand alone)	PERSON 文末記号	
	(Medial)	…, PERSON , …	
話題		上記に当てはまらないPERSON	第三者

4.4.　直接話法の文の主語の選択

4.4.1.　発話中の人名の抽出

図 2 の下線部のように発話中に対話の相手の名前が出現している場合は, 発話中の人名と発話者記述を同定しなければならない．そこで，次の 3 つのステップにより，対話テキスト中に出現する人名や役割名がどの対話参加者に該当するかを推定する．

1. Part-of-speech Tagging, Lemmatizing
2. 固有表現解析
3. パターンマッチによる人名の抽出

1，2 は CoreNLP を利用する．3 については各発話に対して表 1 に示すパターンを上から順に適用する. パターンは形態素の原形の並びで表す.「PERSON」の部分は，固有表現タグが「人」である固有表現,「敬称(Mr., Mrs.など)+固有名詞」, 役割を表す名詞のいずれかと照合する.「文末記号」は「.」や「?」などの文末を表す記号である.「…」は，その他の単語を表す. 相手に呼びかけている時の人名の出現位置は，神谷[5]の中で引用されている B.Douglas[6:1108-1113]によると,「Final」と「Initial」,「Stand alone」,「Medial」の 4 つがあるということから, これらの 4 つのパターンを考えた.「話題」のパターンは, 対話参加者の名前の抽出には不要であるが, 人名を選択する際の制約として用いるため抽出する.

4.4.2.　主語の選択

元の対話テキストの「発話者記述」が「役割名」か「変数名」の場合に以下の方法で発話者記述，すなわち，各発話を直接話法にした場合の主語を決定する.

まず，4.4.1 によって発話者の名前が推定できる場合にはこの人名を用いる．例えば，図 2 の場合，最初の発話(1)と「呼びかけ(Final)」が照合に成功することから，Jane が次の発話(2)の発話者ということが分かるので，このことから B は Jane と推定できる．4.4.1 を使って推定できなかった発話者については, 後で述べる名前辞書から任意の名前を選択する. その際, 対話テキスト中に第三者(4.4.1.

の「話題」で抽出)が出現している場合にはこの第三者とは逆の性別の名前を選択する．これは，she などの三人称代名詞の共参照先の誤りを防ぐためである．名前の性別はファーストネームの場合は後述の名前辞書により，ラストネームの場合は，敬称から判断する．

なお，発話者記述が役割名の場合，人名(固有名)に変換するのではなく，定冠詞 the を付与すること(例:Secretary →The Secretary)で人物(個体)と解釈することが期待できるので，この変換も試みる．

4.4.3. 発話の調整

発話から抽出した PERSON 部分が役割名(例:Waiter)で，かつ，対話参加者のいずれかに対応する場合，発話中の当該役割名を対応する発話者の人名(固有名)に変更する．これにより，共参照の関係づけが可能となる．

4.5. 辞書の構築

4.5.1. 役割辞書

「役割辞書」は，1987～2013 年（奇数年のみ）の間で出題された対話文問題において，発話者記述として出現した役割名(例:Waiter など)と発話中の対話の相手を指しているとわかる役割名(例:発話者記述 Mother=発話中 Mom)を人手で抽出し，構築した．

4.5.2. 名前辞書

「名前辞書」は，1987～2013 年（奇数年のみ）の間で出現した名前（ファーストネームとラストネーム）を人手で抽出し，key を名前，value を性別とし，構築した．性別は，男性を 0 とし，女性を 1，ラストネームの性別は不明であるので，2 としている．性別の判定は，アメリカの人名ランキング[7]より順位の高いほうの性別とした．

5. 実験

共参照解析精度の評価指標として，MUC, B^3, CEAF(e), BLANC[8]を用いる．共参照の正解データは，人手で作成した．各実験で用いたデータを表 2 に示す．「セット 1」は，2003 年以降の対話テキストで構成され，発話者記述は，人名か役割名のみである．「セット 2」は，発話者記述が変数名となるケースで評価するため 2001 年以前の古い出題形式の対話テキストを使用する．「セット 1(役割のみ)」は，「セット 1」のうち役割名となる発話者記述が含まれる対話テキストのみで構成される．

表 2　使用した正解データ

データ名	使用範囲
セット 1	2003～2013 年(奇数年のみ）本試験と追試験(2011 年は本試験のみ)(対話文問題のうち選択肢が文となるもの全 32 問)
セット 1 (役割のみ)	セット 1 中の役割名となる発話者記述を含む問題のみ(6 問)
セット 2	1987～2001 年(奇数年のみ)本試験と追試験(対話文問題のうち選択肢が文となるもの全 66 問)

5.1. 直接話法変換の評価

「セット 1」を用いて，対話テキストをそのまま解析した場合(「A.書き換えなし」と表記)と直接話法の文に書き換えて解析した場合(「B.直接話法変換のみ」と表記)の結果を比較することで，直接話法変換の有効性を評価する．表 3 に結果を示す．直接話法の文に変換することで全ての指標で F 値が 20 ポイント以上向上した．

5.2. 役割名に対する処理の評価

「セット 1（役割のみ）」を用いて，役割名に対して定冠詞を付与する場合（「C.定冠詞付与+B」と表記）と人名に変換する場合（「D.人名変換+B」と表記）の結果を比較することで，役割名に対する処理を決定する．結果を表 4 に示す．人名に変換した方が良い精度が得られることが分かった．

表 3　直接話法変換の評価(P:適合率，R:再現率，F:F 値)

処理	MUC			B^3			CEAF(e)			BLANC		
	P	R	F	P	R	F	P	R	F	P	R	F
A.書き換えなし	57.42	40.98	47.83	55.48	33.8	42.01	45.52	50.73	47.98	46.23	29.8	35.89
B.直接話法変換のみ	78.38	75.61	76.97	74.69	71.18	72.89	68.39	75.52	71.78	68.92	66.5	67.65

表 4　役割名に対する処理の評価(P:適合率，R:再現率，F:F 値)

処理	MUC			B^3			CEAF(e)			BLANC		
	P	R	F	P	R	F	P	R	F	P	R	F
B.直接話法変換のみ	71.42	59.32	64.81	68.66	55.6	61.44	65.86	65.86	65.86	64.65	48.49	55.35
C.定冠詞付与+B	82.14	77.96	80	79.57	74.91	77.17	75.65	79.26	77.41	76.19	72.03	73.94
D.人名変換+B	83.63	77.96	80.7	81.22	74.29	77.6	76.36	76.36	76.36	77.54	70.91	73.98

表 5　変数名に対する処理の評価(P:適合率，R:再現率，F:F 値)

処理	MUC			B^3			CEAF(e)			BLANC		
	P	R	F	P	R	F	P	R	F	P	R	F
B.直接話法変換のみ	33.02	25.78	28.95	35.33	23.31	28.09	34.06	38.38	36.09	23.76	17.66	20.07
E.主語の選択+B	79.48	82.16	80.8	77.2	77.59	77.39	69.46	85.73	76.74	68.05	74.62	70.53

表 6　提案手法の評価(P:適合率，R:再現率，F:F 値)

処理	MUC			B^3			CEAF(e)			BLANC		
	P	R	F	P	R	F	P	R	F	P	R	F
B.直接話法変換のみ	53.76	45.98	49.57	52.26	42.37	46.8	46.88	52.41	49.49	48.47	40.79	44.12
F.統合	80.3	81.8	81.05	77.66	77.24	77.45	70.22	82.93	76.05	69.88	73.76	71.49

5.3.　変数名に対する処理の評価

「セット 2」を用いて，パターンマッチによる発話中の名前の抽出と名前辞書を用いた人名変換(「E.主語の選択+B」と表記)の評価を行う．結果を表 5 に示す．人名に変換することで大幅に精度を向上させることができた．

5.4.　提案手法の評価

「セット 1」と「セット 2」を合わせたデータを用いて，全ての問題に対応した提案手法(「F.統合」と表記)の評価を行う．結果を表 6 に示す．各指標において F 値が約 26～31 ポイント向上した．このことから，共参照解析器の本来の性能を得ることができた．

6.　考察

CoreNLP は，固有表現解析や構文解析などの結果を使用しているため，これらの精度を向上させることで共参照解析の精度を向上させることができると考えられる．「this is 名詞句，人名」のような形の場合，「呼びかけ」のパターンによって抽出される波線部で，構文解析の失敗がみられた．以下の文の場合，下線部の「watch」と「Masako」が同格の修飾語句の関係となっていたため，1 つの mention となり，不要な mention が生成されていた．この mention は，正しい entity クラスタに入っているが不要な mention のため精度を下げる原因の一つとなっている．「Masako」と呼びかけている(4.4.1 の「呼びかけ(Final)」に一致する)箇所は，文の意味と関係がないため，削除することで，不要な mention の生成を防げる．

(削除前)

[Tom]1 said, "[That]3's [[a fancy watch]3, [Masako]2]3.　Was [it]3 expensive?"

[Masako]2 said, "[It]3 sure was.　…"

(削除後)

[Tom]1 said, "[That]2's [a fancy watch]2.　Was [it]2 expensive?"

[Masako]3 said, "[It]2 sure was.　…"

(2005 年センター試験追試験第 2 問 B 問 3)

7.　おわりに

本研究では，既存の共参照解析器で対話形式のテキストを対話形式ではない普通のテキストに自動で書き換える手法を提案し，評価した．その結果，解析器の本来の性能を得ることができた．

今後の方向性として，対話における文の結束性を評価するために，共参照解析がどのように関係するかを調査し，対話文問題の解答手法へと応用したいと考えている．

謝辞

本研究を遂行するにあたり「『ロボットは東大に入れるか』大学入試センター試験関連オンラインタスクデータ」を利用した．データをご提供頂いた「独立法人大学入試センター」および「株式会社ジェイシー教育研究所」に謝意を表する．また，本研究の一部は以下のメンバー(組織)との共同研究として行われた．東中竜一郎，杉山弘晃(以上 NTT)，磯崎秀樹(岡山県立大)，堂坂浩二(秋田県立大)，平博順(大阪工業大)，南泰浩(電気通信大)．記して感謝する．

参考文献

[1] S.Pradhan et al，"CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes"，In proc.CoNLL-2011, 2011

[2] H.Lee et al，"Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task." In proc.CoNLL-2011 Shared Task，2011

[3] 新井紀子ら「ロボットは東大に入れるか？-国立情報学研究所「人工頭脳」プロジェクト-」『人口知能学会誌』27 巻 5 号，pp. 463-469，2012

[4] 松尾文子「英語と日本語の話法」梅光女学院大学英米文学会，英米文学研究，32 巻，pp. 113-130，1996

[5] 神谷健一，「会話における名前の付加-日英対照研究-」『大阪工業大学紀要　人文社会篇』50 巻 1 号，2005 年 10 月，pp43-53

[6] B.Douglas et al，"Longman Grammar of Spoken and Written English　Longman"，1999

[7] http://names.mongabay.com

[8] M.Recasens et al，"BLANC:Implementing the Rand index for coreference evaluation"，Natural Language Engineering Vol.17 Issue 4，pp-485-510，2010