日本語対話の省略のアノテーション

小斉平 ひな（指導教員：戸次大介）

1　はじめに

近年では、スマートスピーカーなどの対話システムが社会に普及しており、人間と計算機の直感的なインターフェイスとして、今後更なるニーズの高まりが期待されている。自然言語処理分野でも対話研究が盛んになっている。

人間の対話には省略が多く現れるという特徴があり、日本語の書き言葉には必須格名詞句の省略が見られることが知られている。日本語の対話では、必須格名詞句の省略に加え、述語の省略やその他文脈から明らかな内容をもつものの省略が頻繁に起こる。

人間同士の対話では、省略されている要素を補って理解することができるが、計算機では、省略要素を補うことが難しく、既存のシステムを適用することができない。

ccg2lambda [1] のような言語理論に基づく意味理解システムは統語解析から始まるが、統語解析は入力が完全な文であることを前提としており、対話のような省略の多い文の処理は得意ではない。そこで、省略をうまく復元することができれば、既存の自然言語理解システムを、対話にも適用することができ、省略を含む対話の意味処理を正しく行うことができると考えられる。

2　先行研究

対話における省略要素のアノテーションの先行研究として、Quan+[2], Quan+[3] による中国語の研究が挙げられる。中国語の省略を含む対話に対して、省略された語・句・節等を復元した完全な文をアノテーションしている。Quan+[2], Quan+[3] は以下のような手法でアノテーションを行っている。クラウドワーカーであるアノテーターが対話を一文ずつ読み、省略のある文を検出する。省略がある文に対して、アノテーターはその発話を、対話の文脈に応じて、省略を復元した完全な文に書き換える。省略のない文に対しては、元の文を維持する。

日本語の省略アノテーションコーパスとして、書き言葉に現れる必須格名詞句の省略を扱った京都大学テキストコーパス [8] がある。京都大学テキストコーパスは、述語に対して省略されている必須格名詞句を復元したものである。このように、日本語の書き言葉に対して現れる一部の省略の要素を対象にした研究は多くあるが、日本語の対話に対して名詞句の省略や主題の省略、述語の省略、格助詞の省略、提題助詞の省略、非飽和名詞の意味的な項を補う省略、概言のムードに先行する要素の省略、照応といった、日本語対話における複数の省略要素を対象とした研究はなされていない。

本研究では、名詞句以外の省略も対象とするため、Quan+[2], Quan+[3] によるアノテーション手法を参考にする。

3　アノテーション

本研究では日本語対話の省略アノテーションの設計とコーパス構築することを目指す。これを実現するために、省略がある対話データに対して、省略された語・句・節、等を復元した完全な文にすることに加え、省略の分類を定義し、分類ラベルをアノテーションする。アノテーションをする対話データにはビジネス対話対訳コーパス [4] を用いる。省略の分類は、小規模のアノテーションを繰り返し、改良を進めていく。本論文では、現在の省略の分類の仕様とアノテーションガイドラインについて説明する。

4　省略の分類

日本語の省略される要素は、名詞句、述語、助詞、等、非常に多岐にわたる [7]。本研究では、省略のタイプを必須格の省略、非飽和名詞の意味的な項を補う省略、助詞の省略、疑問文における述語の省略、概言のムードに先行する要素の省略に分類する。甲斐 [5](pp.2-3) と益岡・田窪 [7](pp.170-171) と西山 [9](pp.33-39) の文献を参考に分類した。

4.1　必須格名詞句の省略の分類

必須格名詞句の省略には統語的省略、談話・状況的省略、場面依存的省略の 3 種類の分類があり、これらは以下のように説明できる。統語的省略とは，複文構造において，生成文法では削除・コピー分析が適用されるような名詞句の省略である。談話・状況的省略は、発話に含まれた内容や発話が行われている場面の要素、その時点で聞き手も共通の話題としている要素で、省略を許される或いは省略をしなければならないタイプのものである [5]。場面依存的省略とは、甲斐 [5](pp.2-3) に依れば、「発話が行われている場面を考慮することによって初めて発話の解釈が可能となるもの」(p.3) である。本研究では、必須格名詞句の省略において、格助詞である「が」「を」「に」「と」の省略のみを省略と認めることにする。統語的省略、談話・状況的省略、場面依存的省略の下位分類として、ガ格・ヲ格・ニ格・ト格を設けた。以下、例 (1b) は例 (1a) の乙の発話にあるニ格必須格名詞句の談話・状況的省略を復元した文である。

(1)　a. 甲：田中さんに会いましたか？
　　乙：ええ、会いました。
　　b. 乙：ええ、田中さんに会いました。

4.2　疑問文における述語の省略

疑問文における述語の省略とは、省略されている要素が文脈から明らかである場合、主題を残して、述語部分を省略することである [7]。以下、例 (2b) は例 (2a) の乙の発話に現れる省略を復元した文である。例 (2b) において復元された乙の「聞いたの」は『疑問文における述語の省略』と分類される。

(2)　a. 甲：ねえ矢澤さん、もう聞いた？
　　乙：何を？

b. 乙：何を聞いたの？

4.3 助詞の省略

助詞の省略には、格助詞の省略と提題助詞の省略がある。

格助詞の省略とは、話し言葉において、発話が行われる場面や文脈から格関係が明らかである場合、格助詞「を」・「が」が省略されることである [7]。

提題助詞の省略とは、提題の「は」が話し言葉で多くの場合省略されることである [7]。

以下、例 (3b) は例 (3a) の格助詞の省略を補った文で、例 (4b) は例 (4a) の提題助詞の省略を補った文である。

(3) a. その本取って。
b. その本を取って。

(4) a. これ、おいくらですか？
b. これは、おいくらですか？

4.4 非飽和名詞の意味的な項を補う省略

飽和名詞は単独で意味的に充足するものであるのに対し、非飽和名詞はそれだけでは意味的に充足せず、パラメータ X を補うことによって意味的に充足するものである [9]。「NP_1 の NP_2」の NP_1 が NP_2 の意味的な項となっている場合、NP_2 を非飽和名詞に分類する。非飽和名詞である NP_2 に対して、NP_2 の意味的な項である NP_1 を文脈や場面に応じて補う。非飽和名詞によっては、複数の意味的な項を補う必要がある場合がある。

以下に非飽和名詞の意味的な項を補う省略の例を示す。

(5) a. ウェイン、調子はどうです?
b. ウェイン、あなたの調子はどうです?

(6) a. (前に A 社の施設の建物が全焼したという対話があって)
だから市場への影響はないと私は思う。
b. だからA 社の施設の建物が全焼したことの市場への影響はないと私は思う。

例 (5b) は例 (5a) に現れる非飽和名詞である「調子」の意味的な項を補った文である。例 (6b) は例 (6a) に現れる非飽和名詞である「影響」の意味的な項を補った文である。非飽和名詞「影響」はサ変動詞由来であり、「X が Y に影響する」のように 2 つの項をとるため、例 (6a) で省略されている X の部分を補う。

4.5 概言のムードに先行する要素の省略

概言のムードに先行する要素の省略とは、相手の言葉に対して、自分が真であると断定できない知識や意見を述べる場合に、「だろう」、「らしい」、「かもしれない」のような概言のムードを表す表現だけを残して、先行する要素を省略することである [7]。

4.6 照応

照応とは、ある表現（代名詞など）が先行文脈によって導入された要素を指示する現象であり、要素を導入する表現を先行詞という。日本語の場合は述語の格要素の位置に出現している代名詞が頻繁に省略されるが、代名詞が省略されているとは考えずに、具体的な表現が省略されていると考えるのでこの場合は「照応」とは扱わず、「名詞句の省略」とする。代名詞が明示的に表れている場合を「照応」と認める。

以下、例 (7b) は例 (7a) の乙の発話にある代名詞「彼女」を、先行詞「花子」に置き換えた文である。

(7) a. 甲：花子はダンスが趣味です。
乙：彼女はダンサーですか？
b. 乙：花子はダンサーですか？

5 現状

日本語の対話に現れる省略の分類とアノテーションガイドラインを作成した。現状として、対話データに対してアノテーションする際に、飽和名詞であるか非飽和名詞であるか、または必須格であるか非必須格であるかの判断が難しい場合がある。以下に、実際にアノテーションをした際に飽和名詞であるか非飽和名詞であるか判断が難しいと感じた例を示す。

(8) 景気はどうですか？

より信頼性のあるアノテーションを行うためにアノテーションガイドラインに言語学的テスト [6] を提示することを考えている。言語学的テストは、日本語話者が持つ言語直観をうまく利用することで、言語学の知識がない人でも正しい判断を行えるように工夫されたものである。

6 おわりに

本研究では、日本語の対話データに対して、省略を復元した完全な文の意味合成を得るために、省略がある文に対して、省略された要素を復元した完全な文をアノテーションし、省略の分類を定義した。

今後の課題としては、アノテーションガイドラインに言語学的テストを提示する等の更なる改良を加え、ガイドラインに従ってコーパスを構築していく。

参考文献

[1] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: a compositional semantics system. In *Proc. of ACL System Demonstrations*, pp. 85–90, 2016.

[2] Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue.

[3] Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling.

[4] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pp. 54–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] 甲斐ますみ. 日本語の省略現象. PhD thesis, 大阪外国語大学, 1999.

[6] 川添愛, 田中リベカ, 戸次大介. MCN コーパス：モダリティ関連表現の曖昧性解消のためのアノテーションと言語学的テストの利用. 2012.

[7] 益岡 隆志・田窪行則. 基礎日本語文法 -改訂版-. くろしお出版, 1992.

[8] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, 1997.

[9] 西山佑司. 日本語名詞句の意味論と語用論 -指示 的名詞句と非指示的名詞句—. ひつじ書房, 2003.