マルチモーダル情報を用いたコンテクスト予測に基づいて伝達文字量を最適化する対話システム

Dialogue system for optimizing the amount of transmitted characters based on context prediction using multimodal information

奥岡 耕平 [1] * 大澤 正彦 [1,2] 今井 倫太 [1]
Kohei Okuoka[1], Masahiko Osawa [1,2], Michita Imai[1]

[1] 慶應義塾大学理工学部
[1] Faculty of Science and Technology, Keio University
[2] 日本学術振興会 特別研究員 (DC1)
[2] Japan Society for the Promotion of Science, Research Fellow (DC1)

Abstract: 人間の対話では省略を用いて伝達文字量の冗長化を防ぐ最適化が行われる。従来研究ではテキストコーパスの文章に対して省略生成を行うことは検討されていたが、対話においてシステムが伝達する文字量を最適化することのユーザーへの影響についての検討は不十分である。そこで本研究では、画像と文章のマルチモーダル情報を用いたコンテクスト推定に基づいて伝達文字量を最適化する対話システムを作成し実験を行い、考察する。

1 はじめに

近年、対話システムに関する研究が盛んに行われており、人間との円滑なコミュニケーションを実現するためのアプローチとして人間の文章に見られる特徴を機械の発話する文章に適用しコミュニケーションの円滑化を図る方法が検討されている。

人間の文章に見られる特徴の 1 つとして、伝達文字量の最適化が挙げられる。人間同士の対話では、省略や指示語を用いて削減し文の冗長度を下げる、伝達文字量の最適化が行われている。特に日本語においては他の言語に比べて省略が頻繁に使われており、省略は日本語を用いた対話システムにおいて重要な最適化の手段である。

従来研究では、文章に対して省略による最適化を行う様々な手法が提案されている。飯田らの研究 [1] では、Nariyama ら [2] の提案した談話的特徴を用いて文書中の語句に対して省略するか否かを分類する自動省略モデルを提案している。しかし、テキストコーパスに対する省略を用いた最適化の精度を向上することに注目しており、対話において省略を行ったときの対話相手への影響は十分に検討されていない。

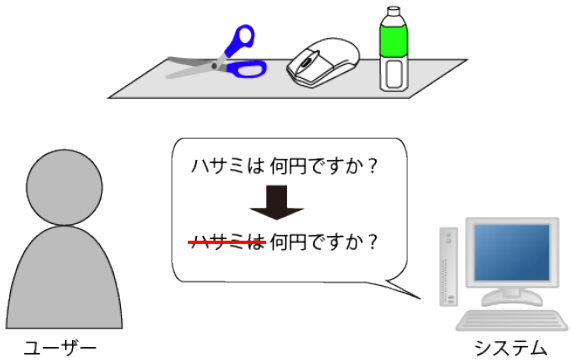

図 1 伝達文字量を最適化する対話システム

そこで本研究では、省略による最適化を行うシステムを作成し、省略を用いた最適化によるユーザーの対話システムの印象への影響を調査した。また、システムの実装において省略の可否を分類する指標として対話におけるコンテクストに着目し、画像と文章のマルチモーダル情報を利用したコンテクスト推定に基づくことで簡易に省略の可否を分類する手法を用いた。作成したシステムを用いて、実験参加者に対話タスクを行ってもらい実験

* 連絡先： 慶應義塾大学理工学部
神奈川県横浜市港北区日吉 3-14-1 26-203
E-mail: okuoka@ailab.ics.keio.ac.jp

参加者へのアンケートの結果から伝達文字量の最適化の有無を被験者間要因として比較し、評価した。

本論文の構成は以下の通りである。まず、第 2 章では本研究で扱う伝達文字量の最適化について解説し、第 3 章では提案するシステムの構成を説明する。第 4 章では実験について解説し、結果を示したうえで結果に対する考察を行う。最後に第 5 章でまとめを述べる。

2 関連研究

人間が文章を生成する際には指示語や代名詞、あるいはゼロ照応と呼ばれる省略などの照応表現を用いることで、文章の冗長度を下げる行為が頻に見られる。

ここで、文章の冗長度を下げる最適化の対話における影響を考えると、最適化は対話における協調の原則である Grice の公準 [3] の示す量の公準を満たすための行為であると考えられる。量の公準は対話において会話の状態に即して相手に情報を過不足なく話すべきというルールで、対話の状況に基づいて冗長度を下げる最適化は対話システムにおいても円滑なコミュニケーションを実現するために考慮する必要があると言える。特に、日本語は他の言語と比べゼロ照応を使うことが多く他言語より省略の重要性は高い。

日本語を対象にした省略生成の研究では滑川ら [4] や Nariyama ら [2] がヒューリスティックな規則に基づく省略生成の手法を提案し、高い精度を示している。また、飯田らの研究 [1] では、Nariyama ら [2] の提案した談話的特徴を用いて文書中の語句に対して省略するか否かを自動的に分類する手法を提案している。

これらの省略生成の手法は効果的であるが、人手による規則の設計や複雑な談話的特徴を扱う必要があるため、本研究では対話における会話の状態を示すコンテクストに着目し、画像と文章のマルチモーダル情報を用いて推定したコンテクストに基づくことで、簡易に省略の可否を分類できるシステムを作成した。

3 伝達文字量を最適化する対話システム

本研究で用いる伝達文字量を最適化する対話システムは図 2 に示す概略図のように 3 つのモジュールから構成されている。以下、各モジュールの内容について述べる。

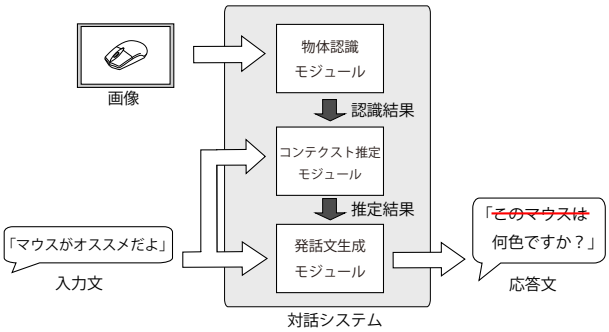

図 2　システム構成図

3.1 物体認識モジュール

物体認識モジュールでは、Web カメラから取得した画像に対して物体認識を行い、認識結果をコンテクスト推定モジュールに伝える処理を行う。本研究では、Tensorflow Object Detection API[5] から、画像データセット Microsoft COCO を用いて学習したモデル「SSD with Inception V2」を使用し、Web カメラで取得した画像に対して物体認識を行い、認識したオブジェクト名をコンテクスト推定モジュールに伝達する。

3.2 コンテクスト推定モジュール

コンテクスト推定モジュールでは、対話相手の発話文に対応するコンテクストを推定し、前状態のコンテクストと比較することで現在の状態に対するコンテクストを決定し、発話文生成モジュールに伝える処理を行う。

まず、物体認識モジュールが認識したオブジェクト名をコンテクスト候補の集合 C として取得する。次に、対話相手の発話文に対して MeCab[*1] を用いて分かち書きを行い、区切られた単語群から品詞が名詞、動詞、形容詞に分類される単語の集合 $\boldsymbol{w}$ を取り出す。続いて、公開されている Word2Vec[6] による学習済みモデル[*2] を用いて単語のベクトル表現を獲得し、事前に決めた閾値を超えるコサイン類似度を用いて各コンテクスト候補 c_j と発話文の関連度を以下のように求める。

$$f(w) = \begin{cases} 0 & (cos(v_{c_j}, v_w) < threshold) \\ cos(v_{c_j}, v_w) & (cos(v_{c_j}, v_w) \geq threshold) \end{cases}$$

$$Similar_{c_j} = \sum_{n}^{\boldsymbol{w}} f(w)$$

[*1] http://taku910.github.io/mecab/

[*2] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

求めた各コンテクスト候補との関連度の中から最大値を事前に決めた閾値と比較し、閾値を上回る関連度を示すコンテクスト候補を発話文に対するコンテクストとする。最大値が閾値を下回る場合、コンテクストは変化しないものとする。

最後に、発話文を受け取る前のコンテクストと推定されたコンテクストを比較し、コンテクストが変化していれば更新し、変化していなければ以前のコンテクストを継続することで現在のコンテクストを決定する。そしてその結果を発話文生成モジュールに伝達する。

3.3 発話文生成モジュール

発話文生成モジュールでは、対話相手の発話に対する応答文を生成すると共に、コンテクスト推定モジュールが推定したコンテクストを基に応答文に対して省略を用いた最適化を行い、対話相手に出力する処理を行う。

応答文の生成にはルールベースによって応答文を選択する手法を用いた。本論文では実験で扱う対話タスクを基にルールを作成し、対話相手の発話に対応する応答文をルールに基づいて出力する。この時、コンテクスト推定モジュールが推定したコンテクストの単語が応答文に含まれている場合はその単語を省略して出力する。

4 実験

4.1 実験方法

提案するシステムの性能を評価するために実験参加者とシステムの対話による実験を行った。ここで、提案システムによる最適化を評価するためには、省略が可能な部分と省略を行うべきでない部分がシステムとの対話中に現れる必要がある。

そこで、本実験では評価に必要な状況が対話中に必ず現れるように、タスクベースの対話実験を行った。実験に用いたタスクは表 1 に示すシナリオに従って、システムが Web カメラで認識している物体について質問を参加者に行い、最終的に参加者にシステムが望む物体を答えてもらうというものである。省略が可能な部分はシナリオの (b),(c),(e),(f),(g) に相当し、シナリオの (h),(i) が省略を行うべきではない部分に相当する。

実験参加者は男性１２名女性４名の平均年齢２３．４歳の計１６名であり、半数の８名の参加者は「最適化を行う」条件、残り半数は「最適化を行わない」条件で実験を行った。実験終了後に、参加者は表 2 に示す７段階のリッカート尺度を用いた全５問のアンケートに答えた。

表 1　対話タスクのシナリオ

	内容
(a)	カメラに映るオブジェクトから 1 つ選択してもらう
(b)	オブジェクトの重さを尋ねる
(c)	オブジェクトの色を尋ねる
(d)	さらにもう 1 つオブジェクトを選択してもらう
(e)	オブジェクトの重さを尋ねる
(f)	オブジェクトの色を尋ねる
(g)	オブジェクトの値段を尋ねる
(h)	初めに選択したオブジェクトの値段を尋ねる
(i)	初めに選択したオブジェクトへの要望を伝える

表 2　質問リスト

	質問内容
Q1	会話中に冗長と感じた部分はありましたか？（1：全くなかった〜7：多々あった）
Q2	会話が成立していないと感じた、または混乱した部分はありましたか？（1：全くなかった〜7：多々あった）
Q3	対話相手に好意的な印象を感じましたか？（1：全く感じなかった〜7：とても感じた）
Q4	対話相手を機械的に感じましたか？、人間的に感じましたか？（1：機械的〜7：人間的）
Q5	システムへの満足度はどのくらいですか？（1：とても不満〜7：とても満足）

4.2 実験結果

実験を行ったところ、作成したルールにマッチしない発話が行われたことでシナリオの進行ができずに対話が破たんしてしまう場合が各条件に 1 度生じてしまった。今回は省略の有無による影響に注目するため、ルールにマッチしない発話によって対話が破たんした２名の参加者のアンケート結果を除いて分析を行った。省略による伝達文字量の最適化の有無を条件に比較したアンケートの結果を図 3 に示す。

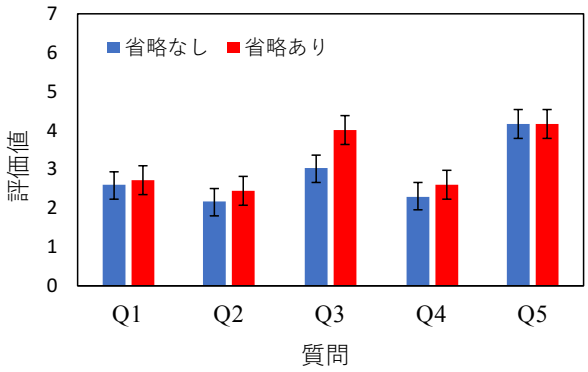

図 3　最適化の有無による比較

分析の結果、すべての質問に有意差は認められないもののシステムを好意的に感じたかという質問については最適化を行った場合の方の平均が上回っており、省略による最適化によってユーザーのシステムへの印象を向上し得ることが示唆された。

また、省略を行った条件の対話ログから省略が可能な部分と省略を行うべきでない部分において適切な省略の可否の選択が行えているかを分析した結果、全ての対話において適切に選択が行われていた。

対して、参加者が正しくコンテクストを理解できたかどうかという点では、理解できずに望む回答が得られなかった場合が 1 名あったものの、その他の参加者は理解できていたため、本実験で使用したシナリオにおいてコンテクスト推定に基づいた省略による最適化はおおよそ有用であることが示された。

4.3　実験考察

4.2 章に示す今回の実験結果からは、会話中の発話文における冗長性に有意差が認められなかった。これは、省略の対象としてコンテクストとしての単語のみであったため、元々の文字量が比較的短いために省略による効果を感じにくかったことが考えられる。また、チャットベースの対話では一度に発話内容が表示されるために、語句の発音にある程度の時間がかかる音声対話と異なり、同じ語句を繰り返すことに対して冗長性を感じにくいと考えられる。そのため、特に音声対話のようなマルチモーダル情報を用いた対話システムにおいては、伝達文字量の最適化を考慮する必要があると考えられる。

会話の成立性においては、有意差が生じなかった。この結果から、提案システムによる省略は応答文に含まれる必要な情報を損なうことなく行われたと考えられる。

しかし、省略を行った文に対してコンテクストを読み取れなかった参加者もいたため、コンテクスト推定の手法や省略以外の最適化手法の検討が必要である。

また、提案するシステムの最適化によってシステムへの印象を向上させ得ることが示された。これは、省略を用いることでシステムの発話する文章が人間の発話する文章に類似し、人間的に感じさせることで印象を向上させたと考えられる。

5　まとめ

本研究では、人間同士の対話に見られる語句の省略等による伝達文字量の最適化をコンテクスト推定に基づいて行うシステムを作成し、ユーザーとのタスクベースの対話による実験を行い、評価した。実験の結果、コンテクスト推定に基づく提案システムは、対話相手に伝えるべき情報を損なうことがないように最適化を行えることが示されると共に、ユーザーのシステムに対する印象を向上させ得ることが示唆された。

参考文献

[1] 飯田龍, 徳永健伸ほか. 日本語書き言葉を対象とした参照表現の自動省略-人間と機械処理の省略傾向の比較. 研究報告音声言語情報処理 (SLP), Vol. 2012, No. 15, pp. 1–10, 2012.

[2] Shigeko Nariyama. Grammar for ellipsis resolution in japanese. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 135–145, 2002.

[3] H Paul Grice, Peter Cole, Jerry Morgan, et al. Logic and conversation. *1975*, pp. 41–58, 1975.

[4] 滑川裕樹, 乾健太郎, 徳永健伸, 田中穂積. 照応・省略を含む日本語論説文生成. 言語処理学会第 5 回年次大会発表論文集, pp. 153–156, 1999.

[5] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*, 2016.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.