トピック情報を用いた口語的記述の省略語推定

原田 智彦　　津田 和彦
筑波大学 システム情報工学研究科 リスク工学専攻
{s1230165@u, tsuda@gssm.otsuka}.tsukuba.ac.jp

1　はじめに

近年, 企業や自治体, 官公庁などが提供する製品やサービスについて, その内容や使い方に関する問い合わせをウェブや電子メール (以下「メール」と略す) を使って行う機会が増えている. 企業など回答を行う側にとっては, 寄せられる問い合わせに対して, 迅速かつ適切な対応を行うことは消費者や住民の信頼を得る重要な要因であり, 反対に怠れば信頼を失いかねない. 不適切な対応がもとでインターネット上に評判が広がり思わぬ対策が必要になる事例も発生しており, 慎重な対応が求められている.

通常, ウェブやメールの問い合わせは, プレーン・テキスト形式であり, 装飾などもない. 最近では携帯電話やスマートフォンが普及し, キーによる文字入力や画面の狭小性のため, 長文を省略する傾向がある. また, 匿名性の高さや比較的若い世代の利用者が多いため, 手紙や報告書などで通常使われる文語調の文体ではなく, より口語調に近いくだけた表現が多くなってきている. 一方で, 対話しながら質問の意図を確認し回答の軌道修正ができる電話と違って, メールによる回答は 1 度返信してしまうと取り返しがつかない. そのため, このような問い合わせテキストの中から意図を正確に読み取ることが重要である.

本稿では, ウェブやメールによる問い合わせテキストの正確な意図理解の支援を目的とし, これらの問い合わせテキストの特徴である口語調による省略に焦点を当て, トピック情報を用いた省略語推定の方法を検討し, 評価実験によって示された効果について報告する.

2　問い合わせテキスト

本稿が扱う問い合わせテキストの例を図 1 に示した. また, 図 2 には本稿で解決しようとする最終的な省略を補完した結果を例示した. 図 1 と図 2 を比べると, ハ格 (主題) やガ格 (主語), ニ格 (間接目的語), ヲ格 (直接目的語) といった文の意味を特徴付ける表層格の多くが省略されていることが分かる. この傾向を大局的に捉えるため, 事前調査を行い, 口語調の表現が多く含まれる「Yahoo!知恵袋」の質問テキスト (16,589 件)[1]と, 口語調でないテキストとして「Wikipedia」の日本語データ[2] を用いて, それぞれ同数に含まれる 4 種類の表層格の 1 文当たりの出現数を集計した. 結果を表 1 に示す. 表 1 から,Yahoo!知恵袋データは Wikipedia に比べて表層格の出現数が約 50%にとどまっており, このことは対象の問い合わせテキストにおける省略語の多さを示している.

パソコンに DVD を入れてもディスクを読み取ってもらえなく、ドライブを見ても常に中身は空です。
しかし CD を再生することは可能です。
こういう場合は修理に出したほうがいいのでしょうか?
すみませんが、ご回答よろしくお願いします。

図 1: 問い合わせテキストの例

(私が) パソコン (のドライブ) に DVD(のメディア) を入れても、(ドライブは) ディスク (の中のデータ) を読み取らず、(私が)(エクスプローラで) ドライブ (の中のファイル) を見ても常に (ドライブの) 中は (ファイルがなく) 空の状態です。
しかし,(私は)(この) パソコンで CD(の音楽や動画) を再生することは可能です。
こういう場合は (私は)(このパソコンを) 修理に出したほうがいいのでしょうか？
すみませんが、ご回答 (を) よろしくお願いします。

図 2: いくつかの省略語を補完した例

表 1: 1 文当たりの表層格の出現数

データ	ハ	ガ	ニ	ヲ
Yahoo!知恵袋 (口語調)	0.274	0.347	0.263	0.385
Wikipdia (非口語調)	0.496	0.604	0.745	0.720

本稿の構成は,3 章で関連する照応と省略の解析に関する研究を紹介し,4 章では省略語推定の方法および評価実験とその結果について報告する.

[1]Yahoo!知恵袋データ 2009 年 3 月分のサブカテゴリー「パソコン」に含まれる全データを使用した.
[2]Wikipedia 日本語データ 2012 年 8 月分の先頭から 16,598 件を使用した.

3 照応と省略の解析

省略された語を補完する処理に照応解析がある. 照応とは, 文中の語句 (先行詞) と語句 (照応詞) が同じ内容を指すことである. 特に日本語は主語の省略 (ゼロ代名詞による照応) が多い言語であるため, 照応解析は自然言語処理の重要なタスクである.

3.1 ゼロ照応解析

照応解析に関する研究は盛んに行われてきたが, その多くは文脈中の表層に現れる手がかりによって解決するものである. 代表的な研究にセンタリング理論 [2] がある. センタリング理論は話題が文脈中の「センター」に維持され, センターが照応や省略の先行詞になりやすいと仮定した考え方である. このセンタリング理論に基づく日本語の照応解析を扱った研究に Walker らの研究 [5] がある.Walker らは先行詞の候補を「主題＞主語＞間接目的語＞直接目的語＞その他」のように顕現性の高いものから並べ, その序列からゼロ代名詞の先行詞を決定する方法を提案している. しかし, センタリング理論では文中に複数の照応詞が存在する場合や前文に先行詞がない場合に適切な解析を行うことができない.

照応解析に関する近年の研究に林部らの研究 [6] がある. 林部らは, センタリング理論に基づく Walker らの方法では,「X を逮捕した」という文のみを手がかりに同じ文脈にある「自首した」のガ格項が X であると判定できないケースに着目し,「格助詞+動詞」を格構造と定義して格構造の類似度と述語項構造解析の履歴を用いる方法を提案し, 従来方法より精度精度が向上することを示した. しかし, この方法では履歴中に先行詞がない場合に適切な解析を行うことができない.

3.2 換喩表現の解析

換喩の現象を扱うことは照応解析においても重要である. 換喩とは比喩の一種であり, ある事物をそれと関連する別の事物に置き換えて表現する現象である. 例えば「漱石を読む」について「漱石」は「漱石の小説」を指していると考えられ, 同様に「電源を入れる」においては「電源」は「電源のスイッチ」を指していると考えられる. 換喩表現は, 図 1 や図 2 の中にも多く現れている. 例えば, 先頭行にある「パソコンに DVD を入れる」では「パソコン」は「パソコンのドライブ」を指し,「DVD を入れても」においては「DVD」は「DVD のメディア」を指している.

また, 換喩表現は係り受け解析の失敗を引き起こすことが知られている. 例えば, 換喩表現である「電源を入れる」では「電源」は「入れる」に係るが, 一方で換喩解釈表現である「電源のスイッチを入れる」では「電源」は「スイッチ」に係り,「スイッチ」は「入れる」に係る. 清田ら [7] は, この換喩表現と換喩解釈表現の係り受け関係のずれが, テキストベースの質問応答システムにおける検索文と対象テキストとのマッチングに影響を与えることから, これを解消するため, 自動抽出した換喩表現を用いた係り受け関係のずれを解消する方法を提案した. 清田らの方法は, 直接的に省略語の解決を目的としたものではないが, 文脈を超えて先行詞の候補を見つけることができている. しかし, どの候補が適切かといった優先度について扱っていない.

表 2 には,1 章で紹介した事前調査において集計した抽出項の文脈[3]内での再現数を示した. これは文中に表層格を抽出した際, 抽出した格の項と同じ語が文脈内に出現した数をデータ 1 件当たりで平均したものである.

表 2: 文脈内での抽出項の再現数

データ	前方文脈	後方文脈	文脈全体
Yahoo!知恵袋 (口語調)	0.244	0.257	0.501
Wikipdia (非口語調)	0.400	0.410	0.810

表 2 から,Yahoo!知恵袋データでは抽出項の再現数が文脈全体でも約 50%にとどまっている.Wikipedia との比較から, 口語調の問い合わせテキストにおける文脈内解決の難しさを示している.

ゼロ照応解析に関する従来の研究は, 口語調でない整ったテキストを対象にしているが, 今後, 口語調を含むテキストに対処するためには, 文脈全体のどこにも先行詞が存在しない場合の解析も必要である.

4 省略語の推定方法

3 章で述べたように従来の方法では同じ文脈中に先行詞がない場合に適切な解析を行うことが難しい. 一般に, 文脈内に候補を持たない省略語を推定するためには探索範囲を文脈外に広げる必要がある. しかし, 探索範囲を広げると, 仮に解決したい省略語の候補集合が予め分かっている場合であっても, 候補語の数が多くなり, 選択はより難しくなる. そこで本稿では, 予め候補集合が与えられた状況下で, 個々の候補語の出現確率を一様ではなく, 文脈に応じて変化すると考え, 言語モデルの一つであるトピックモデルを用いることで大域情報を利用した候補語選択の方法を検討する. なお, 候補集合の収集には 3 章で紹介した清田らの研究で提案されていた方法を用いることとし, 省略語の種類や候補集合の収集方法については本稿では扱わない.

[3]Yahoo!知恵袋データでは 1 件の問い合わせデータを 1 つの文脈と捉え,Wikipedia 日本語データでは 1 段落を 1 つの文脈と捉えて集計した.

本章では, 検討した省略語の選択方法について述べた後, 人工データを用いた評価実験とその結果を報告する.

4.1 トピックモデル

大規模かつ不均質な大量のテキスト情報から, 知識を獲得するための統計的モデリング方法の一つとして, 近年, トピックモデルが注目されている. トピックモデル [3] の基本的な考え方は, 各文書は複数のトピック情報の混合確率分布で表され各トピック情報は単語の確率分布で表されるというものである. 本稿では, トピックモデルとして良い性能を示すことが知られている潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation)[1] を用いる. LDA では, 単語 w の集合を V とし, 単語 $w(\in V)$ の列によって表現された文書の集合とトピック数 K を入力として, 各トピック $z_k(k = 1, ..., K)$ における単語 w の確率分布 $P(w|z_k)(w \in V)$ および各文書 d におけるトピック z_k の確率分布 $P(z_k|d)(k = 1, ..., K)$ を推定する.

LDA を用いることで, 省略を解決する候補語の出現確率を一様ではなく, 文脈に応じて変化すると考え, 単語 $w(\in V)$ の列によって表現された問い合わせテキストの集合 (大域情報) とトピック数 K を入力として, 各トピック $z_k(k = 1, ..., K)$ における単語 w の確率分布 $P(w|z_k)(w \in V)$ および各文書 d におけるトピック z_k の確率分布 $P(z_k|d)(k = 1, ..., K)$ を推定することができる.

4.2 LDA を用いた候補語の選択

前節の方法で推定した LDA モデルを使って, 省略を含む問い合わせテキスト d を入力とし, 与えられた I 個の候補語の集合 C の中から候補語 $w_i(w_i \in C, i, ...I)$ を選択する 3 つの方法を検討した.

1. 単語の出現確率を使った選択 (LDA 単独)
2. 単語の出現確率を使った選択 (N-gram+LDA)
3. トピックの混合確率による分類を使った選択

方法 1 は, 入力した問い合わせテキスト d の中での候補語 w_i の出現確率 $P_{lda}(w_i|d)$ を LDA モデルを用いて推定し, 最も尤度の高い候補語 w_i を選択する.

方法 2 は, 方法 1 と同じ LDA モデルを用いて推定した $P_{lda}(w_i|d)$ と N-gram モデル[4]を用いて推定した $P_{ngram}(w_i|d)$ を次式の線形補完によりスコア化し, 最もスコアの高い候補語 w_i を選択する.

$$score_{w_i}(d) = \lambda P_{ngram}(w_i|d) + (1 - \lambda)P_{lda}(w_i|d) \quad (1)$$

[4]代表的な言語モデルで「ある時点での単語の生起確率は, その直前の個の単語にのみ依存する」と仮定し, 単語の生起を N 重マルコフ課程で近似したモデル

方法 3 は, 機械学習を用いる. 予め訓練用に省略を含む問い合わせテキスト d 'と省略を解決する正解語 $w_i(\in C)$ の組みを用意し, 正解語 w_i をクラスとして問い合わせテキスト d' を学習する. このときの素性には LDA モデルで推定した d '中のトピックの混合確率 $P(z_k|d$ ' $)$ を用いる. 選択時には, 入力した問い合わせテキスト d に対して LDA モデルで推定した d 中のトピックの混合確率 $P(z_k|d)$ を使い, 機械学習によって分類されたクラス w_i を選択する.

4.3 人工データによる評価実験

評価実験には,3 章で紹介した清田らの研究で提案されていた換喩表現ペアの抽出方法を利用した. これは換喩表現および換喩解釈表現の出現パターンに基づいたパターンマッチングを用いる方法である. これにより実験データに対して予め換喩表現-解釈表現のペアを作成し, 得られたペアの中から, 換喩解釈表現に存在し換喩解釈表現に存在しない語の集合を抽出し, これを候補語集合 C とする.

実験は, テストデータ中に換喩解釈表現パターンとマッチする表現が見つかった場合に, マッチした部分を換喩表現に置換し, 置換によって欠落した語を正解とする方法で行う.

実験データには,Yahoo!知恵袋データ (収集期間 2004/4/1∼2009/4/7) の質問データ 3,206,559 件を使用した. その中からサブカテゴリー「パソコン」のデータ 576,841 件を抽出し, このうち 500,000 件を 100,000 件ずつに 5 分割し, うち 4 つを LDA モデルの推定用に, 残り 1 つを評価用に使用して交差検定を行う.

実験に使用する換喩表現と候補語集合は, 同じ 576,841 件のデータの中から, 換喩表現-解釈表現ペア 38,538 件[5]を作成し, その中から換喩解釈表現の出現数合計が上位のペアのうち, 候補語の出現傾向の違いから今回は次の 2 つの表現を選び実験対象とした.

1. 「元の【候補語】に戻す」(候補語 134 件)
2. 「パソコンの【候補語】について」(候補語 225 件)

表現 1 は候補語ごとの出現数に偏りがあり, 表現 2 は候補語が均等に出現する傾向がある.

なお,LDA の実装には GibbsLDA++[4] を用いた. LDA モデルの推定に必要なトピック数 K やディリクレ分布のハイパーパラメータ α, β は, 予め実験データの一部を使ったパープレキシティ測定により, モデルの精度の良いトピック数 $K = 150$ とハイパーパラメータ $\alpha = 0.1, \beta = 0.01$ に決定した[6]. 方法 2 の N-gram

[5]換喩表現 1,259,662 件と換喩解釈表現 167,816 件から換喩表現-換喩解釈表現ペアを 38,538 件作成した.

[6]交差検定用に 5 分割した 1 つを使用し,$K = \{30, 50, 150, 200\}$, $\alpha = \{0.01, 0.1, 0.5, 1, 1.5\}, \beta = \{0.01, 0.1, 0.5, 1, 1.5\}$ の組み合わせでパープレキシティを測定した.

モデルの実装にはSRILMツールキットを用いた[7]. 式(1)の線形補完係数λは事前実験で高い精度を示した値を使用し, 表現1は$\lambda = 0.85$, 表現2は$\lambda = 0.5$とした. 方法3の機械学習にはデータマイニングソフトウェアWEKAを利用した[8].

4.4 結果と考察

実験データ500,000件から表現1の「元の【候補語】に戻す」が776件, 表現2の「パソコンの【候補語】について」が2,294件見つかった. 表3および4に, それぞれの正解率を示した. なお, 比較のため, 各表には最も出現頻度の高い候補語(最頻語)を正解とした場合の正解率とN-gram単独で推定した正解率も合わせて表示した.

表 3: 表現1の推定結果

構成比	最頻語	N-gram	方法1	方法2	方法3
正解	25.26%	38.02%	12.76%	40.59%	42.00%
不正解	74.74%	61.98%	87.24%	59.41%	58.00%

表 4: 表現2の推定結果

構成比	最頻語	N-gram	方法1	方法2	方法3
正解	7.02%	11.38%	26.68%	33.35%	34.00%
不正解	92.98%	88.62%	73.32%	66.65%	66.00%

表3では,LDA単独で利用した方法1は最頻語やN-gram単独の場合に比べて正解率が低く,N-gramとLDAを組み合わせた方法2は最頻語やN-gram単独の場合よりも僅かではあるが正解率が高い. また機械学習を利用した方法3は最も正解率が高い. 表4では,LDA単独で利用した方法1は最頻語やN-gram単独の場合に比べて正解率が高く,N-gramとLDAを組み合わせた方法2はさらに正解率が高い. 機械学習を利用した方法3は最も正解率が高いことが分かる. 表3, 表4を比較すると,N-gram単独では, 候補語ごとの出現数に偏りがある表現1より, 候補語が均等に出現する表現2の方が正解率が低い. 一方でLDA単独で利用した方法1では, 候補語ごとの出現数に偏りがある表現1より, 候補語が均等に出現する表現2の方が正解率が高いことが分かる.

表3, 表4の結果から, 方法2が,LDAと線形補完することでN-gramモデルで解消できない候補語選択の曖昧性を一部解消し, 正解率を改善したことを確認することができる. また, 方法3については, 各候補語にトピック情報の混合確率を学習させることにより, 候補語とトピック情報の不一致を修正し, 候補語とトピックを強く結びつける効果があったと考えられる. また, 表3, 表4の比較から, 方法1が, 候補語ごとの出現数に偏りがない場合にN-gramモデルよりも候補語の選択に良く機能することを確認することができる.

これらの結果から,LDAモデルを利用するにより, 従来の言語モデルであるN-gramモデルを補完して候補語の選択性能を向上させる効果やトピック情報と候補語を機械学習により結びつける分類で選択性能を向上させる効果を確認した.

5 おわりに

本稿では, ウェブやメールによる問い合わせテキストの正確な意図理解の支援を目的とし, 問い合わせテキストの特徴である口語調による省略に焦点を当て, トピック情報を用いた省略語推定の方法を検討し, 評価実験の結果を示した.

実験の結果, 予め候補語集合が与えられた省略語の選択において, トピック情報を用いることで, 従来の言語モデルであるN-gramモデルを補完して候補語の選択性能を向上させる効果やトピック情報と候補語を機械学習により結びつけることで選択性能を向上させる効果があることを確認した.

謝辞

本研究の実施にあたっては, ヤフー株式会社が国立情報学研究所に提供した「Yahoo!知恵袋データ (第2版) を利用した.

参考文献

[1] D M Blei, A Y Ng, and M I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

[2] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, June 1995.

[3] Thomas Hofmann. Probabilistic latent semantic indexing. In *the 22nd annual international ACM SIGIR conference*, pp. 50–57, New York, New York, USA, 1999. ACM Press.

[4] Xuan-Hieu Phan and Cam-Tu Nguyen. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). 2007.

[5] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese Discourse and the Process of Centering. *arXiv.org*, Vol. cmp-lg, , September 1996.

[6] 林部祐太, 小町守, 松本裕治. 文脈情報と格構造の類似度を用いた日本語文間述語項構造解析. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2011, No. 10, pp. 1–8, May 2011.

[7] 清田陽司, 黒橋禎夫, 木戸冬子. 自動抽出した換喩表現を用いた係り受け関係のずれの解消. 自然言語処理, Vol. 11, No. 4, October 2004.

[7] モデルは5gram確率を用い, パラメータには補完モデルにinterpolate, スムージングにkndiscountを使用した.

[8] 事前実験による比較検討から最も高い識別性能を示したLogitBoostを選択した. 他のオプションについてはデフォルトのままとした.