E-049

記述位置情報による質問構造のモデリング手法

Method to modelling question structure by a described position

原田智彦 †　　津田和彦 †
Tomohiko HARADA　　Kazuhiko TSUDA

1 はじめに

企業や自治体, 官公庁にとって, 寄せられる消費者や住民からの問い合わせに迅速かつ適切に回答をすることは, 信頼性を得るための重要な要因である. 近年, この問い合わせを電子メール (以下「メール」と略す) で行うことが増加している.

通常, これらメールによる問い合わせは, プレーン・テキスト形式である. そのため, 文字装飾などもなく, テキスト中に記述された重要な質問を見落としたり, 質問の意図を取り違えて的外れな回答をするなどの問題が発生している. また, メールの場合, 電話での対応のように, リアルタイムでの対話によって問い合わせ内容を確認したり, 追加情報を収集することが難しい. その結果, 顧客満足度の低下や問題解決の長期化につながる事例も発生している.

このような課題に対し, 本研究では, 自然言語処理を利用してテキスト中に記述された質問文の正確な理解を支援することを目指しいる. 本稿では,Q&A サイトや特定のシステムなどに関するユーザーからの質問の主要部分が, 質問と付加説明によって構成されていることに着目する. その上で, 質問者が最も聞きたい事柄が質問中のどこに書かれているのか, という記述位置情報から分析を行い, テキスト中に含まれる質問構造をモデリングする手法について述べる.

2 文書構造の分析

問い合わせメールのテキスト中には, 質問と質問に至った背景や具体的な事象を説明する付加説明によって構成される. テキストには明確なフォーマットや記述ルールがなく, 書き手や読み手のスキルに依存するため, 質問文の文書構造を正確に理解することが難しい. ここでは, 文書構造に注目した分析を行った先行研究を紹介する.

長谷川ら [1] は, メールから重要文を抽出し, 携帯電話向けの要約システムへの適用を提案している. まずメールの文書構造を「先頭文」「冒頭文」「通常文」「引用」「署名」に区別する. 次に本文中の各文に「名乗り」「依頼」「疑問」「話題」「スケジュールの通知」を特定する形態素パターンとそれに対応するスコアからなるヒューリスティックなルールを適用し, ランキング順で重要文を抽出している. 主たる目的が要約であることから, 本研究のように質問文や付加説明文の関係性に立ち入っていないが, 本研究では提案された文書構造の整理手法を参考にした.

高木ら [2] は, 類似文書検索において, 検索質問文書に複数の主題があることに注目し, 主題ごとに類文書検索を行い, 主題ごとの主題重要度を用いて検索結果を生成する方法を提案している. 主題の抽出にはテキストセグメンテーションやパターンマッチングを利用している. 主たる目的が特許検索であることから, 特許文書の構造が前提となり, そのまま問い合わせメールのような自由記述に適用させることは難しい.

田村ら [3] は, 従来の質問応答システム研究では扱っていない複数文質問の質問タイプを同定する方法を提案している. まず, 与えられた複数文質問の中から質問タイプを決める際に最も重要な 1 文を核文として抽出し, 核文かどうかの判定には単語の情報を素性として用いた分類器を使っている. 複数文質問の質問タイプは, 核文となる 1 文で同定できるという仮定を導入し, 複数文質問の中にある複数の質問文を扱っていない点が本研究と異なる.

以上のように, 文書構造に注目した分析を行った研究はいくつか存在するが, 複数の質問文を対象として, 質問文と付加説明文の関係性に注目している研究は少ない.

3 質問構造モデリングの手順

以下では, 本手法により, テキスト中の質問文や付加説明文の記述位置情報から質問構造をモデリングする手順を示す.

step1　文ごとへの分割
　パターンマッチングを使い「。」や「.」の句点,「！」や「？」[*1]の記号および行末を分割点として, テキストを複数の文に分割する.

step2　文の役割の付与
　パターンマッチングを使い,「ますか」「ませんか」「ですか」「でしょうか」などの質問表現を含む文を「質問」に区分し, その他の文は「説明」に区分する. また, 文末に「おしえて」「おねがい」[*2]などの表現を含む文を新たに「その他 (挨拶や依頼文など)」に区分する.

step3　記述位置が連続する文のグループ化
　次に,「説明」に区分された文同士や「その他 (挨拶や依頼文など)」についても同様に, 記述位置が連続する同じ役割の文をひとつのグループにまとめる.

step4　質問構造のパターン収集
　step1〜3 により,「説明」→「質問」→「説明」→「質問」→「その他 (挨拶や依頼文など)」の役割の並びによる質問構造のパターンが得られる. この手順を対象の質問データに対して繰り返し, 質問構造のパターンを収集する.

step5　テンプレートの当てはめ
　step4 で収集した質問構造のパターンに対して, 図 1 のようなテンプレートの当てはめを検討する.

† 筑波大学

[*1] 「？」は口語記述の中では, 付加疑問的に使用されるケースが多いため, 助詞「か」と連続して出現する「か？」を分割点とした.

[*2] 「おねがい」を含む文でも, 文中に「ヲ格」(対象格) が見つかれば「質問」に区分した.

図 1 は, 質問の数=2 の場合の典型的な質問構造のパターンを示したものである. 矢印は「説明」が付加説明している「質問」つまり「係り先」を示す. 本稿では, この質問構造の基本パターンを「質問構造テンプレート」と名づける.

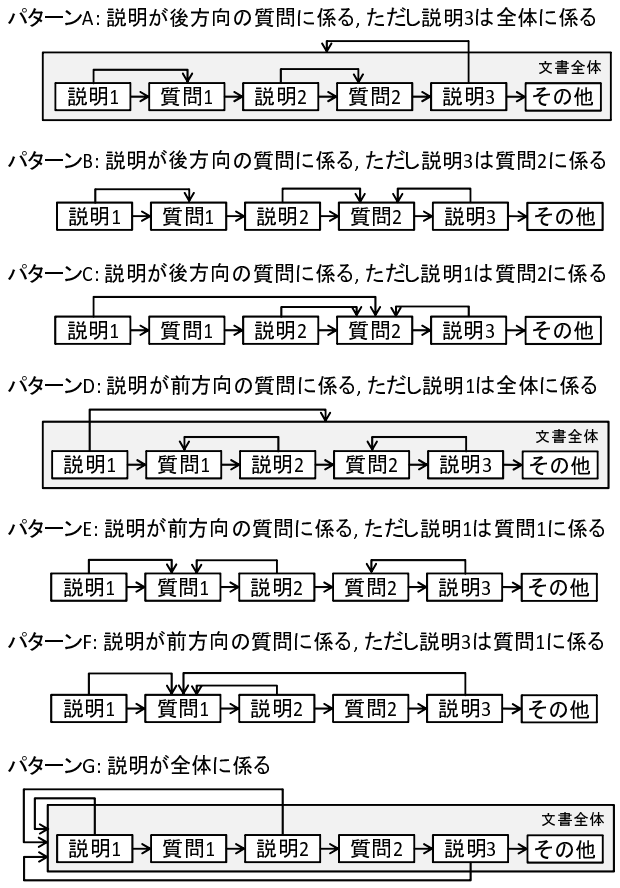

図 1　質問構造テンプレート (質問の数=2 の場合)

4　質問構造モデリング事例

QA サイト[*3]の質問データを利用して, 本手法の適用を行った.

データには Yahoo!知恵袋の質問データを利用した[*4]. この中で「一定のボリュームがある」「背景知識を前提としない内容である」「将来のデータへの適用し易さ」などの観点から,2009 年 3 月分の「パソコン」に関する質問データ 16,599 件を選択した. これを step1 の「文ごとへの分割」により,55,891 文に分割した.

対象の質問データに対して step1〜4 の手順を実施し, 文の役割の並びによる質問構造のパターンを収集した.

表 1 に, 質問データに含まれる質問文の数を示している. 表 1 から, 対象データでは質問文を 3 つ以上含むデータが全体の 4.3% と僅かである. そのため,step5 では質問文の数=2 のデータに着目することとした.

表 2 は, 質問文の数=2 のデータについて質問構造のパターンの数が多い順に並べた上位 10 件と, その全体に対する構成比を示す. 文の役割の並びが, 図 1 の質問構造テンプレートと完全に一致したのは, 表 1 のランク=10(3.85%) である. 他の 9 件も一部の要素を省略したものに過ぎず, テンプレートが当てはまる. なお, 各パターンへの当てはめは今後の課題である.

表 1　質問文の数

質問文の数	相対度数	累積度数
≤ 1	80.98%	80.98%
$= 2$	14.72%	95.70%
≥ 3	4.30%	100.00%

表 2　質問構造のパターン (質問文の数=2 の場合)

ランク	文の役割の並び	相対度数	累積度数
1	質問→質問	26.51%	26.51%
2	説明→質問→質問	15.43%	41.94%
3	説明→質問→質問→その他	8.55%	50.49%
4	説明→質問→質問→説明	6.83%	57.32%
5	質問→質問→その他	5.20%	62.52%
6	説明→質問→質問→説明→その他	4.75%	67.27%
7	質問→質問→説明	4.71%	71.97%
8	説明→質問→説明→質問	4.66%	76.64%
9	質問→説明→質問	4.26%	80.89%
10	説明→質問→説明→質問→その他	3.85%	84.74%

5　おわりに

QA サイトの質問データを使い, テキスト中の質問や説明の記述位置情報から, テキスト中に含まれる質問構造をモデリングする手法について述べた.

本手法を適用し, 質問構造テンプレートを当てはめることで, テキスト中に記述された問い合わせの可視化レベルを高め, 質問内容の取り違えを減らしたり, 質問の数の把握とチェックが容易になることで回答漏れを減らすなど, メール対応の対応品質向上と効率化が期待できる.

今後は, 本稿で扱っていない, 質問文と付加説明文の係り受け関係や質問文と質問文が連続する場合の分析について詳細化していく.

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] 長谷川隆明, 林良彦, & 山崎毅文. (2004). 電子メールにおける重要文抽出と携帯電話向け要約システムへの適用 (コンテンツ処理). 情報処理学会論文誌, 45(7), 1745-1754. 一般社団法人情報処理学会.

[2] 高木徹, 藤井敦, & 石川徹也. (2005). 検索質問の主題分析に基づく類似文書検索と特許検索への応用 (情報検索). 情報処理学会論文誌, 46(4), 1074-1081. 一般社団法人情報処理学会.

[3] 田村晃裕, 高村大也, & 奥村学. (2006). 複数文質問のタイプ同定 (自然言語). 情報処理学会論文誌, 47(6), 1954-1962. 一般社団法人情報処理学会.

[*3] QA サイトとは, 質問を載せておくと, 他のユーザーが答えてくれるサイトであり, 投稿される質問および回答件数は年々増加している.

[*4] 研究者向けに, 期間 2004-04-01〜2009-04-07 の 16,257,413 件の質問データが提供されている.