問い合わせ文章の階層化のための抽出型要約および係り受け解析を用いた重要箇所抽出

林 岳晴　　大段 秀顕　　竹中 一秀　　湯浅 晃　　大木 環美
株式会社 NTT データ
{Takeharu.Hayashi, Hideaki.Odan, Kazuhide.Takenaka, Akira.Yuasa, Megumi.Ohki}@nttdata.com

概要

本稿では，問い合わせ文章からの FAQ 自動抽出を目的とした段階的なクラスタリングの精度向上に寄与する，重要箇所の抽出手法を提案する．提案手法では，FAQ の構造がカテゴリ＞質問核心＞質問条件と段階的な情報からなることに着想を受け，重要箇所を段階的に自動抽出する．実験により，抽出型要約による重要文抽出，および係り受け解析による重要文からの質問核心抽出の各ステップで，8 割以上の性能を達成することを確認した．また，カテゴリ単位での FAQ 自動抽出において，従来手法と比較し精度改善が見られることを確認した．

1 はじめに

企業の問い合わせ対応業務の効率化には「よくある質問（FAQ）」を整備し運用する手段がよくとられるが，大量にある問い合わせから FAQ を作成するには膨大な人件費がかかる．そこで，FAQ 整備のために，問い合わせ文章をクラスタリングし FAQ を自動抽出する手法がよく用いられる．

一般的に，FAQ はコンテンツを階層的にまとめることで，検索しやすくなると言われている．たとえば，カテゴリやサブカテゴリの単位でまとめると，ユーザーは FAQ のカテゴリ検索が可能となる．また，シナリオ型チャットボットにおいては，階層数を 3〜5 程度にすることが良いとされる．階層が深くなりすぎるとユーザーが求める FAQ に辿り着くまでの所要時間が増えてしまうため，階層数を少なく抑えることが良いとされる[1][2].

そのため，カテゴリ＞質問核心＞質問条件と問い合わせを段階的に詳細化できれば，ユーザーは求める答えに辿り着きやすくなり，関連質問を体系的に整理できると考えられる．このように段階的に文章を集約する方法として，階層型クラスタリングが用いられているが，入力として問い合わせ文章の全文を用いるのではなく，カテゴリ・質問核心・質問条件に該当する各部分を用いれば，各階層の集約に不要な情報を排除でき，集約精度の向上が期待できる．

そこで本研究では，問い合わせ文章からカテゴリ・質問核心・質問条件の各部分を自動抽出する手法を提案する．さらに抽出した各部分を入力として段階的なクラスタリング（多段階クラスタリング）を行い，FAQ 自動抽出における精度改善効果を検証する．

2 関連研究

問い合わせの構成要素を抽出するには，文節也文同士の関係性に着目する必要がある．このような文章を構造化する手法には，大別して，文章意味理解を目的としたものと，重要箇所の抽出を目的としたものがある．文章意味理解を目的とした研究として，横山ら(2003)[3]は，文間の関係タイプを因果関係や背景情報など 8 つの関係性で定義し，これらを機械学習手法で自動特定することで文章を構造化する手法を提案している．肥塚ら(2007)[4]は，述語項構造の各項に的確な意味役割を付与することを目的として，SVM や述語との係り受け関係を用いる手法を提案している．一方，重要箇所を抽出する目的においては文書要約に関するタスクとして研究されており，深層学習を用いる手法[5][6]等が近年では数多く報告されている．これらの研究は，文章の構造を汎用的に定義し解析しようとするものであり，FAQ 自動抽出に向けた文章構造化を目的とはしていない．

FAQ 抽出を行う目的においては，壹岐ら(2018)[7]は，教師あり学習による抽出型要約を用いて，不要文の多い問い合わせログから重要箇所を抽出する手法を提案しているが，これは多段階クラスタリングに特化した文章の構造化を目的とはしていない．

そこで本研究では，問い合わせ文章はカテゴリに関する部分（対象部），質問核心に関する部分（核

心部），質問条件など詳細部分（周辺部）から構成されると定義し，問い合わせ文章から各部分を自動抽出する手法を提案する．

3 問い合わせ文章の構造の定義

筆者らによる事前の分析の結果，各問い合わせ文章は 3 つの必要情報で構成されることがわかった．この分析から，問い合わせを以下の構造として定義する．

- 対象部…文章の主題（何について問い合わせているか）を示す部分
- 核心部…文章の核心（問い合わせている内容は何か）を示す部分
- 周辺部…対象部・核心部以外の，背景情報や参考情報に該当する部分

また，対象部および核心部を合わせた部分が問い合わせ文章の重要部分に該当すると考えられる．これを核文とする．

- 核文…文章中の重要部分（対象部＋核心部）

上記の定義をもとに，問い合わせ文章の構造を解析し，多段階クラスタリングの入力とすることで，効果的な FAQ が抽出できると考えた．具体的には，FAQ を抽出する際に，1 階層目では対象部，2 階層目では核心部，3 階層目では周辺部と，多段階クラスタリングを実行することで，カテゴリ＞質問核心＞質問条件の 3 階層ツリー構造として集約できる．

4 部分抽出器の生成

3 章で定義した構造を問い合わせ文章から抽出する流れについて図 1 に示す．

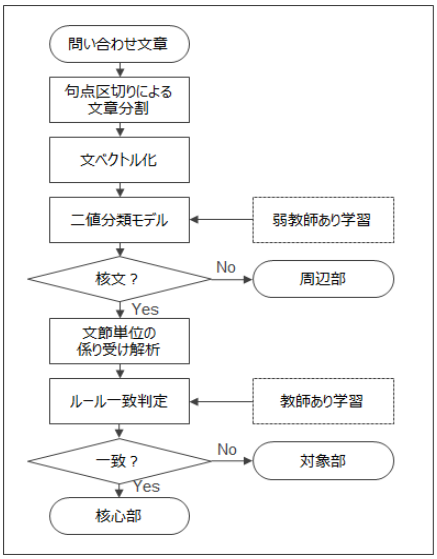

図 1 部分抽出処理の流れ

具体的には，次の 2 ステップの処理を行う．

1. 問い合わせ文章を文単位に分割し，機械学習により，核文として意味的に重要な文を判定する．核文以外の文を周辺部とする
2. 各核文に対し，係り受け解析を用いて核心部を抽出し，核心部以外の文節を対象部とする

4.1 核文・周辺部の抽出器

文章中からの重要文抽出においては，抽出型要約が一般的に用いられ，Yang ら(2019) [8]の BERTSum 等教師あり学習による手法や Kamil ら(2018) [9]の EmbedRank，EmbedRank++等教師なし手法があるが，本手法ではより質問文に特化した石垣ら(2018)[10]の枠組みを参考とする独自手法を用いる．

石垣らは，冗長な質問文を，より短く理解が容易な単一文質問に変換することを目的として，抽出型要約の枠組みを提案している．大量の質問投稿を対象に，1 文で構成される質問を正例，10 文以上の文から構成される質問投稿の各文を負例として，擬似的な教師データを作り，各文を要約に含めるべきか否かを判定する二値分類モデルの学習に用いている．特徴ベクトルの構築には単語および品詞タグの unigram，bigram，trigram を用い，分類器にはロジスティック回帰モデルを用いている．本手法では石垣らの提案した枠組みを用い，OKWave に投稿された IT カテゴリの質問 441,503 件から，正例 61,286 件および負例 305,719 件を擬似的な教師データとして生成した．そして生成した擬似的な教師データを用いて，問い合わせ文章の各文が核文か周辺部かを判定する二値分類モデルを学習する．ただし本手法は，特徴ベクトル生成モデルおよび二値分類モデルに関して石垣らの手法とは異なる．具体的には，以下のモデルを検証し，ベースライン手法と比較する．

- BERT[11]二値分類モデル
- TfidfVectorizer 及び線形回帰モデル
- EmbedRank/EmbedRank++及び線形回帰モデル

ベースライン手法としては，先頭の質問文が最も重要であるとする手法（Lead-Q）[12]を用いる．

4.2 核心部・対象部の抽出器

抽出された核文を入力として，文節単位で核心部か対象部かの判定を行う．日本語は Li ら(1976)[13]の提唱した主題優勢言語にあたり，主題が先で内容が後の構成を取る場合が多い．この特性を利用して，文の末尾から内容に相当する核心部を抽出し，核心

部以外の部分を主題となる対象部として抽出する．さらに，問い合わせ文において核心部に当たる部分の文節間の係り受け構造を見ると，同様のパターンが多く見られる．そのため，核心部として抽出すべき述語との係り受け関係を予め特定することで，核心部抽出のルールとして設定できると考えた．

IT サービスに関する Enterprise の問い合わせ文章データセットから，644 件の問い合わせに対して核心部を人手で分析した結果を表 1 に示す．核心部の係り受け構造として頻出するパターンに傾向があることがわかった．たとえば，主語名詞（nsubj）や副詞修飾語（advmod）が文末の ROOT に直接係る場合に，核心部となることが多い．これらから一部重複や短すぎるパターンを除外して，213 件の係り受け構造パターンを核心部抽出のルールとして設定した．

なお，本手法では係り受け解析器として GiNZA[14]を用いている．

表 1 核心部とすべき係り受け構造例

係り受け構造	出現数	核心部の例
nsubj, ROOT	162	エラーが，表示される．
advmod, ROOT	79	どう，すればよいのでしょうか．
ROOT	53	なぜでしょうか．
obj, ROOT	32	操作手順を，教えてほしい．
advmod, advcl, ROOT	30	どう，すれば，解消されますか．
obl, ROOT	29	エラーで，起動しません．
advcl, ROOT	24	ログインできなく，なりました．
amod, advcl, ROOT	22	どの，ように，すればよいですか．

5 評価および試行実験

4 章で提案した構文解析器を用いた部分抽出と，多段階クラスタリングによる FAQ 自動抽出の実験を行った．

5.1 実験設定

データセットには，4 章で用いた IT サービスに関する問い合わせ文章データセットを用いた．核心部抽出ルールの特定に用いていない 161 件の問い合わせを対象に部分抽出の評価を行った．また，定性分析のために，公開データである NII の Yahoo!知恵袋データ（第 2 版）の質問本文をサンプリングして用い，試行実験を行った．

5.2 核文抽出の評価

表 2 に核文抽出の評価結果を示す．Acc(sent)は文単位での正解率を，Acc(all)は複数文から構成される問い合わせ文章のすべての文に対して正解した割合を示す．

表 2 核文抽出の正解率

	Acc(sent)	Acc(all)
Lead-Q	0.81	0.39
BERT	**0.83**	**0.40**
TF-IDF	0.79	0.29
EmbedRank++	0.61	0.14
EmbedRank	0.59	0.12

BERT で特徴量を抽出したモデルがもっとも高い正解率となり，ベースライン手法（Lead-Q）の精度を上回った．

5.3 核心部抽出の評価

次に，表 3 に核心部抽出の結果を示す．ここで，入力には 5.2 節で抽出した核文を用いた．8 割以上の高い精度を達成し，抽出したルールが妥当であることを示している．

表 3 核文からの核心部抽出の正解率

	Precision	Recall	F1
GiNZA	0.80	0.81	0.81

5.4 部分抽出試行実験

部分抽出の出力例を図 2 に示す. 図 2 の(1)～(3)の抽出結果では，対象部は質問カテゴリ，核心部は質問核心, 周辺部は質問条件にそれぞれ対応しており，問い合わせ文章を想定通りに構造化できている．

一方, 抽出した対象部や核心部から, 質問の話題・内容が読み取れないような誤りもあった. たとえば, 図 2 の(4)の例では「教えて下さい！！」の部分のみ核文と判定されており，主要な箇所を抽出できていない．この例では，問い合わせ文章中に，重要箇所が複数文にまたがっている．核文判定の学習では，1 文からなる質問文のみを正解として学習しているため，直接的な質問内容のみが核文として判定されてしまったと考えられる．

また，図 2 の(5)の例は，核心部抽出の誤り事例で

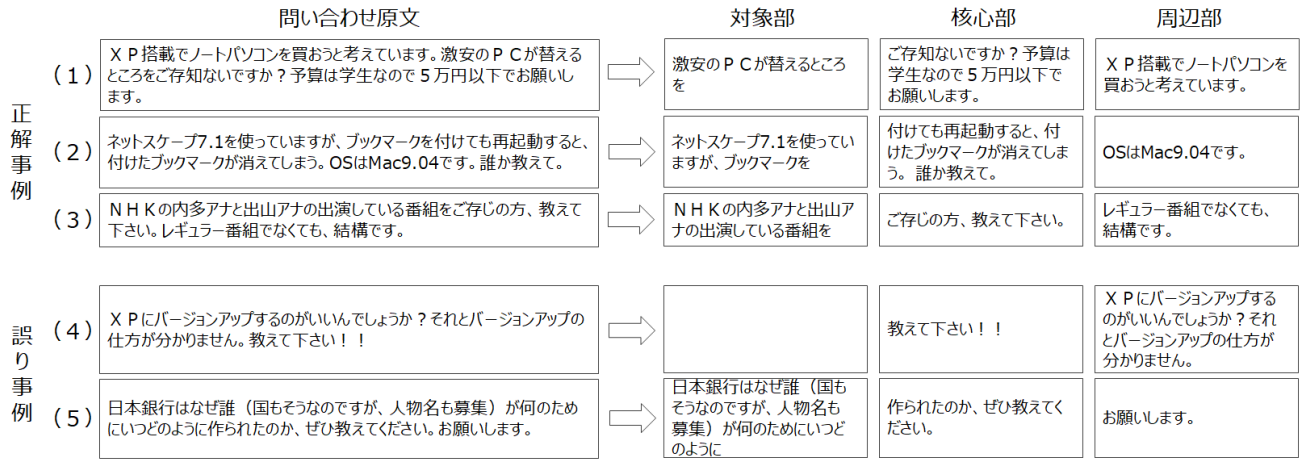

図 2 部分抽出結果例

ある．対象部に含まれている「なぜ～どのように」の部分は質問の内容核心に当たる部分であり，本来核心部に含まれるべきである．この例では，核心部の述部である「作られたのか」に対して，「なぜ」「誰が」「何のために」「いつ」「どのように」の複数の文節が本質的な係り受け関係にある．これらをすべて核心部と判定するためには，核心部の係り受け構造パターンが同一の問い合わせ文章を用いてルール抽出する必要がある．しかし，文の途中で補足情報として「（国もそうなのですが，人物名も募集）」が含まれており，本質的な係り受け関係を解析することは困難である．そのため，問い合わせ文の文整形を行うことや，「誰が」「何のために」等の用語を含む文節を核心部として抽出することで，核心部の抽出性能を向上することができると考えている．

6 多段階クラスタリングへの適用

部分抽出で得られた対象部・核心部・周辺部の各部分を入力とすることで，多段階クラスタリングによる FAQ 自動抽出の精度改善効果を検証した．

4 章および 5 章で用いた IT サービスに関する問い合わせ文章データセットに対して，次の手順で多段階クラスタリングを行った．

1. 各問い合わせ文章の対象部についてクラスタリングを行う
2. 各カテゴリクラスタ内の問い合わせ文章の核心部でクラスタリングを行う

上記の手順により，1 階層目がカテゴリレベル，2 階層目が質問核心レベルとなる FAQ の自動抽出を実現できる．ベースライン手法として，カテゴリ・質問核心の両方の階層において，どちらも問い合わせ文章の全文を入力とした場合の多段階クラスタリングを用いた比較検証結果を表 4 に示す．

カテゴリレベルの集約では，対象部を入力とする場合が文章全体を入力とするベースラインを上回った．一方，質問核心レベルでの集約においては，文章全体を入力とした場合の方が精度が高い結果となった．部分抽出の大きな課題は，5.4 節に示したとおり，核心部抽出において係り受け関係の複雑な文の解析の難しさであるため，文整形やルールベース手法との組み合わせにより精度を改善できると考えている．

表 4 多段階クラスタリングへの適用結果(F1)

	カテゴリ	質問核心
提案手法（対象部・核心部を入力とした多段階クラスタリング）	**0.57**	0.46
ベースライン手法（全文を入力とした多段階クラスタリング）	0.56	**0.55**

7 おわりに

本稿では，抽出型要約による重要文抽出，および係り受け解析による重要文からの質問核心抽出の各ステップで，8 割以上の性能を達成することを確認した．また，部分抽出で得られた対象部・核心部で，多段階クラスタリングによる FAQ 自動抽出を行うと，カテゴリレベルでの集約において若干の精度改善が見られることを確認した．

今後の課題としては，本手法に加えて，「誰」「何のために」等の用語を含む文節を核心部として抽出する等，ルールベース手法を組み合わせることが考えられる．また問い合わせ文章からの FAQ 自動抽出に限らず，文書検索や文書要約にも提案手法を適用することが可能であるか検証したい．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 2 版）」を使用させて頂きました.

参考文献

[1] -, FAQ チャットボットはシナリオが大事｜作り方・ポイントを解説｜トラムシステム (引用日: 2022 年 1 月 12 日.) www.tramsystem.jp/voice/voice-3322/

[2] -, 顧客満足度が上がるチャットボットシナリオの作り方【カスタマーサポート編】(引用日: 2022 年 1 月 12 日.)

[3] 横山憲司, 難波英嗣, 奥村学. Support Vector Machine を用いた談話構造解析. 情報処理学会研究報告 自然言語処理研究会報告, Vol. 2003 No.23, pp. 193–200, 2003.

[4] 肥塚真輔, 岡本紘幸, 斎藤博昭, 小原京子. 日本語フレームネットに基づく意味役割推定. 一般社団法人言語処理学会 自然言語処理, Vol. 2007-14 No.1, pp. 43–66, 2007.

[5] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv: 1912.08777v3 [cs.CL], 2020.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv: 1910.13461v1 [cs.CL], 2019.

[7] 壹岐太一, 田嶋隼平, 下沢将, 比屋根一雄. ヘルプデスクの対応記録からの QA リストの半自動抽出. 研究報告知能システム (ICS), 2018.12: 1-8, 2018. sinclo.medialink-ml.co.jp/blog/chatbot-customer-support-schenario/

[8] Yang Liu. Fine-tune BERT for Extractive Summarization. arXiv: 1903.10318v2 [cs.CL], 2019.

[9] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, Martin Jaggi. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. arXiv: 1801.04470v3 [cs.CL], 2018.

[10] 横山憲司, 難波英嗣, 奥村学. Distant Supervision による質問集約. 情報処理学会研究報告, Vol.2018-NL-236 No.5, pp. 1–5, 2018.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805v2 [cs.CL], 2018.

[12] Tatsuya Ishigaki, Hiroya Takamura, Manabu Okumura. Summarizing Lengthy Questions. Proceedings of IJCNLP2017, Vol.1, pp.792-800, 2017.

[13] Li Charles N., Thompson Sandra A. "Subject and Topic: A New Typology of Language". In Charles N. Li. Subject and Topic. New York: Academic Press. 1976.

[14] 松田寛, 大村舞, 浅原正幸. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会 第 25 回年次大会 発表論文集, 2019.