共起単語ペアの組合せ最適化を用いた問合せログのクラスタリング

湯浅 晃　野村 雄司
株式会社 NTT データ
{Akira.Yuasa, Yuji.Nomura}@nttdata.com

概要

本稿では企業のコンタクトセンタに蓄積される問合せログを集約し「よくある質問」(FAQ)を作成するユースケースに焦点を当てたクラスタリング手法を提案する．提案手法は「文書カテゴリを表現する特徴的な単語の最適な組合せが選択できれば，それらが含まれる文書群を抽出することで良い集約ができるのではないか」との着想をもとに問合せログ集約を最大被覆問題として定式化し，数理最適化により単語ペアの組合せを得る．本実験により，従来手法を上回る FAQ 抽出精度が得られ，また，その他の利点として最適クラスタ数が自動的に得られることと，高い解釈性が得られることを確認した．

1 はじめに

企業のコンタクトセンタやチャットボットにおいて用いられる FAQ の作成は人手で行われることが多く，コスト上の課題である．FAQ の検索タスクに関する研究は数多いが，問合せログからの FAQ 作成に関する研究は比較的少なく，精度に課題があることがわかっている．そこで FAQ を自動抽出するために，問合せログから類似する問合せを集約するタスクに取り組む．

本タスクの特徴は，入力が比較的短い問い合わせ文であり，また出力として実用的な粒度の FAQ を得ることを目的とすることである．ここで実用的な粒度とは，問い合わせ内容を端的に表し，対応する回答が得られる粒度を意味する．例えば EC サイトにおける実用的な粒度の FAQ とは，「注文」などの抽象度の高い表現ではなく，「注文をキャンセルする手順を教えてほしい」といった表現となる．

本研究では，まず人手によって問合せ文書を集約する場合に行われる操作を観察した．その結果，文書中の主題や対象を表す特徴語が抽出され，それらを用いた階層的な分類がなされることに着目した．そこで「文書カテゴリを表現する特徴的な単語の最適な組合せが選択できれば，それらが含まれる文書群を抽出することで良い集約ができるのではないか」との着想をもとに，問合せ文書の集約を単語ペアの組合せ最適化問題として定式化し，ベイズ最適化および線形最適化を用いた手法の検証を行った．

2 関連研究

問い合わせログからの FAQ 抽出の全体的な処理を示した研究として，飯塚ら[1]は，抽出的要約，word2vec による文書ベクトル化，孤立文書の除去，および k-means によるクラスタリングといった一連の手法を示している．個々の処理に関する研究として，友松ら[2]は文書のベクトル化における BERT の使用を提案している．またクラスタリングの代わりにパターン抽出を用いるアプローチとして，長谷川ら[3]は，係り受け関係にある２語の組をパターンとして扱い，応答文書の長さを考慮した上で，頻度ベースの特徴語選択と重要度に基づきパターンを抽出する手法を提案している．

FAQ 抽出の一連のプロセスの中で，本稿ではクラスタリングに焦点を当てて取り組む．なぜなら，抽出的要約や孤立文書の除去といった前段の技術は，ノイズを含むデータセットから真に重要な部分を抽出するためのものであるが，ノイズが含まれない理想的な問い合わせデータセットであっても，実用的な FAQ を得ることを目的とした場合に，そのクラスタリングの精度に課題があることがわかっているためである．これは実験結果を示す表 3 において，ノイズが含まれないデータセットである e-learn, StackFAQ であっても，既存手法の精度が十分でないことによって示している．

短いテキストのクラスタリングにおける主要課題は，単語数が少ないため文書ベクトルが高次元かつスパースとなることである．この課題に対処するため単語埋め込みやテキスト拡張の手法が考案されており，特に深層学習を用いた手法はデータを低次元

でより適切に分離可能な表現空間にマッピングする効果的な方法として近年勢いを増している．深層表現学習とクラスタリングの統合に焦点を当てた研究は多数あり，Xu ら[4]の STCC においては Word2Vec と K-means を組合せた手法が提案され，その後の Hadifar ら[5]の STA においては，SIF を使用することで単語埋め込みを強化した. Rakib ら[6]の ECIC では，凝集クラスタリングを用いた後に，各クラスタに含まれる外れ値を除くサンプルをラベル付きデータとみなす半教師あり学習を反復する手法が用いられている．Zhang ら[7]の SCCL では対照学習が用いられ，同一ラベルのサンプルからデータ拡張されたサンプルをまとめ上げ，異なるラベルのサンプルからそれらを押しのける学習方法がとられている．

次に，頻出アイテムセットに着目するアプローチとして，Fung ら[8]の FIHC 法が代表的である．これは文書集合においてグローバルに頻繁に使用されるアイテムセットごとに，アイテムセットを含むすべての文書で初期クラスタを構築した後に，クラスタ内のアイテムセット頻度と，単一文書内のアイテムセット頻度を考慮し，文書をクラスタに紐付けていく操作を順次行うものである．その後の発展研究として Zhang ら[9]の Maximum Capturing 法, Lee ら[10]の OCFI 法等が挙げられる．

3 提案手法

3.1 手法検討

本研究ではビジネス適用における現実的な制約条件を考慮した上で手法検討を行った．まずベクトル空間アプローチで高精度が得られている Rakib らの ECIC においては反復学習を行うため, サンプル数が全体として大きく，またクラスタサイズに偏りが少ないケースが望ましい．しかし本研究の検証用データセットにおいては，表 1 に示す通り，サンプル数が 500 から 1,000 件程度と比較的小さく，また Enterprise データセットの標準偏差が非常に大きいことからもわかるように，実世界の問い合わせデータにおけるクラスタサイズは概してロングテールの形をとるため適用が難しいと判断した．Zhang らの SCCL についてはまず対照学習のためのデータ拡張を行う必要があり，その精度が問題となったため適用を見送った．

次に, Fung らの頻出アイテムセットアプローチの適用を検討したが，アイテムセットのグローバル頻度と各サンプル文書内における頻度を使用するため，比較的大きな語彙数の文書を扱う必要がある．しかし, 本稿の問題設定においては 1 文書あたり 5-10 語程度と文書内のアイテムセット頻度が十分にとれるほどの長さではないため適さないと判断した．

このように既存手法については，いずれもデータセットの特性と手法のミスマッチが問題となった．そこで，問題を特徴語の組み合わせ最適化問題として捉える手法を提案する．

3.2 組合せ最適化問題としての定式化

提案手法の核となるアイデアは「文書カテゴリを表現する特徴的な単語の最適な組合せが選択できれば，それらが含まれる文書群を抽出することで良い集約ができるのではないか」というものである．この着想に基づき，以下の定式化を行った．

N件の問合せ文書の集合である問合せログ $D = \{d_1, d_2, \dots, d_N\}$が与えられ，$D$に含まれる単語集合 $W = \{w_1, w_2, \dots, w_M\}$ から単語ペアをL個抽出し，$\{(w_{11}, w_{12}), (w_{21}, w_{22}), \dots, ((w_{L1}, w_{L2}))\} \in W$ とする．次に抽出した単語ペアが含まれる文書を D から抽出し, その結果得られた文書集合を$\{C_1, C_2, \dots, C_L\}$とする．このとき，本問題を最大被覆問題として，目的関数と制約条件を以下の通り設定する．

$$\max \quad . \sum_{i=1}^{L} \frac{|C_i|}{N} \tag{1}$$

$$s.t. \quad \bigcap_{i=1}^{n} C_i = \emptyset \tag{2}$$

$$L \leq L_{max} \tag{3}$$

式(1)は目的関数であり，選択された単語ペアを含む文書数を全文書数で割ったものであり，変数は単語ペアの選び方である．式(2)，(3)は制約条件であり，式(2)は各クラスタに含まれる文書集合に重複する文書が存在しないことを, 式(3)は, クラスタ数Lはハイパーパラメータとして与える最大クラスタ数L_{max}以下とすることを表す．

3.3 評価方法

提案手法はクラスタリングを行っているが，本タスクの目的は FAQ 抽出である．そこで評価方法としては purity や AMI 等を用いたクラスタリング自体の評価ではなく，人手で定めた正解 FAQ クラス

タがどの程度正しくクラスタリングの結果抽出できているかを測ることによる評価を行う．

まず各文書に対して人手で文書内容を考慮した上でクラスタリングを行い，各クラスタに一意のラベルを付与しこれを正解ラベルとした．次に，自動抽出された各クラスタに含まれる文書について，クラスタ内の同一ラベルの文書数の割合(純度:purity)が閾値以上を占めるマジョリティとなる正解ラベルを抽出クラスタの推定ラベルとした．抽出されたクラスタにおいて，マジョリティとなる正解ラベルが存在しない場合，推定失敗を意味する推定ラベルを付与した．このようにして得た正解ラベルと推定ラベルを用いて，適合率，再現率，および f1 スコアを算出した．パラメータ設定について，ここでは純度の閾値は 0.5 を設定した．

4 実験設定

データセット　Sumikawa ら[11]の e ラーニングデータセット(e-learn)，StackFAQ[12]データセット(StackFAQ)および，筆者らの所属企業内部におけるコンタクトセンタ応対ログデータセット(Enterprise)，の 3 種類のデータセットを用いた．各データセットの基礎情報を表 1 に示す．

表 1　データセットの基礎情報

データ特性	e-learn	StackFAQ	Enterprise
文書数	427	1249	779
単語数の平均値	4.7	6.2	10.7
単語数の標準偏差	2.5	2.1	5.8
クラスタ数	79	125	60
クラスタサイズの平均値	5.4	10	13
クラスタサイズの標準偏差	1.1	0.1	37.4

前処理　StackFAQ データセットは，Google Cloud Translation API[i]を用いて英語から日本語に翻訳した．全てのデータセットについて，MeCab を用いて形態素解析を実施し，品詞フィルタ処理として名詞，動詞，形容詞のみを残し，原形化を行った．

最適化ツール　線形最適化においては，最適化モデラーとして PuLP[ii]を，ソルバーとして CBC を用いた．ベイズ最適化においては Optuna[iii]を使用し，最適化アルゴリズムは NSGA-Ⅱを用いた．

[i] https://cloud.google.com/translate/docs/reference/rest/v3/projects/translateText
[ii] https://coin-or.github.io/pulp/

5 実験

5.1 予備実験

予備実験では提案手法のポテンシャルを検証するために，FAQ の抽出精度が従来手法と比較して高くなるような単語ペアの組合せが存在するか確認した．

ベースラインとして TFIDF＋HAC(階層的凝集クラスタリング)を用いた．HAC のアルゴリズムは群平均法を，距離尺度はユークリッド距離を用いた．クラスタ数は，正解ラベルが既知の前提とし f1 を最大化するクラスタ数を採用した．またベクトル化手法による比較のために fastText+HAC を検証した．

次に，ベイズ最適化を用いて最適な単語ペアを探索した．このとき正解クラスタラベルが既知の前提で，目的関数を正解 FAQ 抽出の f1 とした．試行回数を重ねた収束時における FAQ 抽出の f1 スコアを表 2 に示す．

表 2　ベースラインおよびベイズ最適化の結果

手法	e-learn	StackFAQ	Enterprise
TFIDF+HAC(最大値)	0.65	0.64	0.46
fastText +HAC(最大値)	0.68	0.62	0.44
BayesOpt	0.91	0.99	0.86

表 2 に示すようにベイズ最適化による FAQ 抽出精度はベースラインの TFIDF+HAC を大きく上回っている．この結果により，もし適切な単語ペアの組み合わせが選択できたとすると，その単語ペアを含む文書の集合をクラスタとして抽出することで，高い FAQ 抽出精度を得られることが確認できた．

5.2 本実験

線形最適化　3.2 に示した通り，文書のクラスタリングを最大被覆問題として定式化し，文書カバレッジを最大化する単語ペアの組み合わせを求めた．単語ペア選定におけるハイパーパラメータとして，最小クラスタサイズを 3 とした．これは 3 個以上の文書に含まれる単語ペアのみが選定対象となることを意味する．最大クラスタ数をパラメータとして増加させたときの結果を図 1 に示す．最適化の各試行においていずれの場合も最適解が得られた．また，クラスタ数は単語ペア数が一定となる点で f1 も最大

[iii] https://optuna.org/

となるため，その時の単語ペア数をクラスタ数として採用すればよいとわかる．

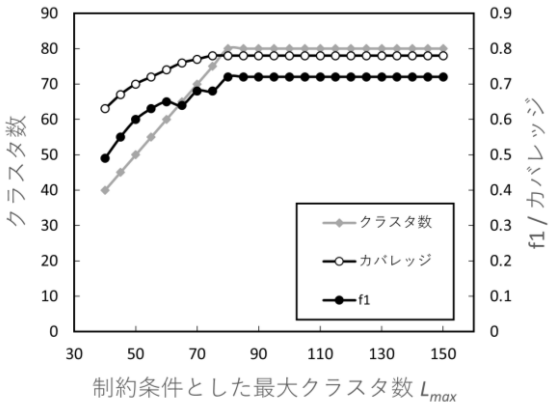

図 1 線形最適化の実行結果(e-learn)

ベイズ最適化　次に，ベイズ最適化手法を用いた実験を行った．予備実験では提案手法のポテンシャルを測るために正解ラベルが既知の前提で，目的関数を FAQ 抽出の f1 としたが，本実験では，実ビジネス利用を想定し，正解ラベルの情報を使用しない以下の目的関数を用いた．

$$\max \quad .w_c \cdot Coverage(C) + w_o \cdot Overlap(C) \qquad (4)$$

$$Coverage(C) = \sum_{i=1}^{L} \frac{|C_i|}{N} \qquad (5)$$

$$Overlap(C) = \left(1 + \alpha \frac{\bigcap_{k=1}^{n} C_k}{\bigcup_{k=1}^{n} C_k}\right)^{-1} \qquad (6)$$

ここで，式(4)の目的関数は，カバレッジスコアとオーバーラップスコアの重み付き和である．式(6)に示すオーバーラップスコアは，各クラスタに共通的に含まれる文書が少ないほど大きい値をとる関数である．w_cおよびw_oは重みであり，$w_c = 1$および$w_o = 2$を用いた．α はハイパーパラメータであり，9 を用いた．結果を図 2 に示す．

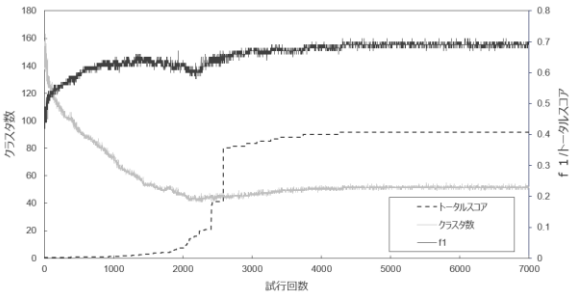

図 2.ベイズ最適化の実行結果(e-learn)

図 2 に示すように，試行回数を増やすにしたがって目的関数であるトータルスコアは増加し，また f1 スコアも増加し，その後一定に収束している．このときクラスタ数も一定に収束しており，最適クラスタ数として得られることがわかる．

最終的な実験結果を表 3 に示す

表 3　線形最適化およびベイズ最適化の精度

手法	e-learn	StackFAQ	Enterprise
TFIDF+HAC(区間平均)	0.62	0.58	0.23
fastText +HAC(区間平均)	0.56	0.55	0.29
LinearOpt	0.72	0.68	0.43
BayesOpt	0.69	0.74	0.47

TFIDF＋HAC および fastText+HAC については，実利用を想定すると最適クラスタ数は未知であることを考慮し，クラスタ数を真のクラスタ数の 75%から 125%までの区間で等間隔に 10 点とり，10 点の f1 値の平均をとってスコアとした．

表 3 に示すように e-learn および Enterprise データセットにおいては，線形最適化およびベイズ最適化を用いた提案手法によりベースラインを上回る精度が得られた．その他の利点として提案手法では両手法ともに最適クラスタ数が得られ，また単語ペアが含まれていることがすなわちそのクラスタの特徴となるため，高い解釈性が得られることがわかった．

6 おわりに

問合せログからの FAQ 抽出を目的として，特徴語の組み合わせ最適化問題として解く手法を提案した．本実験では提案手法により従来手法を上回る精度が得られた．また精度以外の利点として，最適クラスタ数が自動的に得られることと，単語ペアの組み合わせがクラスタの特徴を示すため高いクラスタの解釈性が得られることがわかった．今後の課題として，FAQ 抽出精度を高めるために，単語ペア選択において，問合せ内容と関係のないノイズとなる単語の排除や，同義語の考慮が必要である．またベイズ最適化においては目的関数に凝集性等のクラスタとしての望ましさを考慮した改善が課題となる．

参考文献

[1] 飯塚新司, 菊地大介, 宮内秀彰, 髙橋毅, and 黒澤隆也. "ヘルプデスクの問合せデータを用いた FAQ 抽出技術の研究." *日立ソリューションズ東日本技報= Hitachi Solutions East Japan technical report* 25 (2019): 31-34.

[2] 友松祐太; 戸田隆道; 杉山雅和. AI チャットボットのためのチューニング支援システム. In: 人工知能学会研究会資料 言語・音声理解と対話処理研究会 90 回. 一般社団法人 人工知能学会, 2020. p. 08.

[3] 長谷川友治, et al. コールセンターにおける大規模質問応答データに基づく FAQ 作成支援システムの実装. 第 66 回全国大会講演論文集, 2004, 2004.1: 73-74.

[4] Xu, Jiaming, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. "Self-taught convolutional neural networks for short text clustering." Neural Networks 88 (2017): 22-31.

[5] Hadifar, Amir, Lucas Sterckx, Thomas Demeester, and Chris Develder. "A self-training approach for short text clustering." In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pp. 194-199. 2019.

[6] Rakib, Md Rashadul Hasan, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. "Enhancement of short text clustering by iterative classification." In International Conference on Applications of Natural Language to Information Systems, pp. 105-117. Springer, Cham, 2020.

[7] Zhang, Dejiao, Feng Nan, Xiaokai Wei, Shangwen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew Arnold, and Bing Xiang. "Supporting Clustering with Contrastive Learning." arXiv preprint arXiv:2103.12953 (2021).

[8] Fung, Benjamin CM, Ke Wang, and Martin Ester. "Hierarchical document clustering using frequent itemsets." In Proceedings of the 2003 SIAM international conference on data mining, pp. 59-70. Society for Industrial and Applied Mathematics, 2003.

[9] Zhang, Wen, Taketoshi Yoshida, Xijin Tang, and Qing Wang. "Text clustering using frequent itemsets." Knowledge-Based Systems 23, no. 5 (2010): 379-388.

[10] Lee, Cheng-Jhe, Chiun-Chieh Hsu, and Da-Ren Chen. "A hierarchical document clustering approach with frequent itemsets." International journal of engineering and technology 9, no. 2 (2017): 174.

[11] Sumikawa, Yasunobu, Masaaki Fujiyoshi, Hisashi Hatakeyama, and Masahiro Nagai. "Supporting creation of FAQ dataset for E-learning chatbot." In Intelligent Decision Technologies 2019, pp. 3-13. Springer, Singapore, 2020.

[12] Karan, Mladen, and Jan Šnajder. "Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval." Expert Systems with Applications 91 (2018): 418-433.