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

Extraction of Frequently Asked Questions from Texts in Customer Inquiry Data

ヘルプデスクでは, 問合せ内容に近い質問を高いヒット率で検索できるFAQを作成するため，テキストのクラスタリング技術を用いて蓄積された問合せデータからよくある質問を抽出し，FAQ に登録している。しかし従来手法では，問合せに多く含まれる表現の揺れに対応するため，同義語を定義した辞書の整備が必要となる。そこで本研究では，辞書が不要な分散表現を用いたクラスタリングと，クラスタから代表テキストを自動抽出する手法を考案した。評価の結果，分散表現を用いた FAQ 検索で 90%以上の高いヒット率を実現できることを確認した。一方，クラスタリング結果は，問合せ内容の特定などに課題があり，人手による分類と近い結果が得られない場合があることが分かった。

飯塚　新司　Iizuka Shinji
菊地　大介　Kikuchi Daisuke
宮内　秀彰　Miyauchi Hideaki
髙橋　毅　Takahashi Tsuyoshi
黒澤　隆也　Kurosawa Ryuya

1. はじめに

コールセンターなどのヘルプデスクでは，サービスの品質向上のため，よくある質問とその回答をまとめたFAQを整備している。応対時にヘルプデスクのオペレータはFAQデータベースに対して検索を行い，ユーザからの問合せに対する回答を確認する。このとき，問合せ内容に近い質問が検索でヒットする確率(ヒット率)が高いほど，より迅速に正確な回答をすることができ，サービス向上の効果が期待できる。このため，ヘルプデスクではヒット率の高いFAQデータベースを作成することが求められている。これを実現するには，ユーザからの問合せ頻度が高いよくある質問をFAQデータベースに登録することが望ましい。そこで，顧客管理システムなどに蓄積された問合せデータのテキストから，よくある質問を抽出して，FAQデータベースに追加することが行われている。従来，よくある質問を抽出する作業を支援する技術として，テキストのクラスタリング技術が用いられてきた。しかし，問合せに多く含まれる言い回しの違いや同義語といった表現の揺れに対応するためには，従来手法では同義語を定義した辞書の整備が必要となる。そこで，近年，同義語の辞書が不要な技術として類似文書検索などで利用されるようになった分散表現に着目し，分散表現を用いたFAQ抽出技術を検討した。

2. 従来手法

問合せデータのテキストからよくある質問を抽出する際に用いられている，従来のクラスタリング技術について述べる。テキストのクラスタリングでは，各クラスタ内に類似した内容のテキストが分類される。データ件数の多いクラスタから代表的な質問内容のテキスト(以下，代表テキストと呼ぶ)を適切に抽出することで，よくある質問の抽出が可能となる。クラスタリングには，単語の出現頻度に基づく指標である TF-IDF(Term Frequency - Inverse Document Frequency)を用いてテキストを数値ベクトルに変換し，数値ベクトルに対してクラスタリングを行う手法[1)]や，トピック抽出技術でテキストを分類する手法[2)]が用いられている。

3. 分散表現を用いた FAQ 抽出技術

3.1 分散表現による類似文書検索

分散表現によるテキストのベクトル化[3)]は，テキストを数百次元の数値ベクトルで表現する手法である。TF-IDF によるベクトル化と比べて次元が少ないので，k-means 法のような従来の数値ベクトルに対するクラスタリング技術により適している。また，ベクトルの向きの近さを表す指標であるコサイン距離が，指定した閾値以下となるベクトルをもつテキストを抽出することで，

内容が類似しているテキストを検索することができる。これを，分散表現による類似文書検索と呼ぶ。ヘルプデスクの問合せには，言い回しの違いや同義語による表現の揺れが多く見られる。分散表現による類似文書検索は，同義語の辞書を用いなくてもこのような表現の揺れに対応できるため，問合せに類似する FAQ や応対履歴のレコメンドに用いられている [2]。

3.2 分散表現を用いたFAQ抽出技術の課題

FAQ 検索で分散表現が用いられるようになった一方で，FAQ 抽出技術については，従来手法では分散表現は考慮されていない。そのため，表現の揺れに対応するには同義語の辞書の整備が必要である。その手間を削減するため，辞書が不要な分散表現を用いた FAQ 抽出技術を新たに考案した。FAQ 検索には 3.1 節の分散表現による類似文書検索を用いることを前提とした。

1 章で述べたように，ヘルプデスクではヒット率の高い FAQ データベースが求められる。また，FAQ 抽出におけるテキストのクラスタリングでは，各クラスタ内に人が読んだときに類似した内容のテキストが分類されていることが前提となる。この前提を満たさないと，抽出した代表テキストがクラスタ内の他のテキストと類似せず，FAQ 検索で類似しないテキストがヒットする可能性が生じる。そのため，テキストのクラスタリング結果は人手でテキストを分類したときと近い結果になることが望ましい。以上をまとめると，(1) 分散表現による類似文書検索で高いヒット率を実現できる，(2) テキストのクラスタリング結果が人手でテキストを分類したときと近い結果である，という 2 つの要件を満たす FAQ 抽出技術を考案することが課題となる。

4. 提案手法

提案手法では，分散表現によりテキストを数値ベクトルに変換し，数値ベクトルに対するクラスタリング技術であるk-means法でテキストのクラスタリングを行う。クラスタの中心点に最も近い分散表現を持つテキストをクラスタの代表テキストとする。これにより，分散表現による類似文書検索をFAQ検索に用いた場合に高いヒット率が期待できる。また，人手で作成したクラスタの正解データを用いて最適なパラメータを選択することで，テキストのクラスタリング結果が人手でテキストを分類したときと近い結果になるようにする。下記(1)から(6)で提案手法の処理手順を説明する。

(1) データクレンジングと前処理

データクレンジングおよび前処理として，定型文・固有表現・不要語の除去，形態素解析による単語への分かち書きを行う。必要に応じて，抽出的テキスト要約 [4] で主要な内容を含む重要文を抽出する。

(2) テキストの分散表現の算出

テキストsがl個の単語$w_1\ w_2\ \cdots\ w_l$へと分かち書きされているとき，テキストsの分散表現$x(s)$を，各単語w_tの分散表現$x(w_t)$の合成により算出する。本研究では，単語の分散表現の学習アルゴリズムに word2vec [3] を用いた。以下，(1)の処理で出力されたテキストのうち，空文字列でないものを分析対象のテキストとし$s_1, s_2, \dots, s_N$で表す。また，その分散表現$x(s_i)$を$x_i (i = 1, 2, \dots, N)$と表記する。

(3) 孤立テキストの除去

分析対象のテキストのうち，内容が類似するテキスト(以下，類似テキスト)が一定件数未満のテキストを孤立テキストと呼び，テキストのクラスタリングの対象から除外する。以下，ベクトルxとyのコサイン距離を$\mathrm{Cos}(x, y) = 1 - (x \cdot y)/(\|x\|\|y\|)$とする。テキストの類似の判定におけるコサイン距離の閾値を$d > 0$としたとき，$\mathrm{Cos}(x_i, x_j) \le d$ならば，テキスト$s_i$と$s_j$は類似していると判定する。テキスト$s_i$に対し，類似テキストの件数を$\mathrm{Sim}_d(s_i) = \#\{1 \le j \le N : \mathrm{Cos}(x_i, x_j) \le d\}$とする。ここで，$\#A$は集合$A$の要素数を表す記号である。類似テキストの件数の閾値をnとしたとき，$\mathrm{Sim}_d(s_i) \le n$ならば，テキスト$s_i$を孤立テキストと判定する。

(4) テキストのクラスタリング

孤立テキストを除いた分析対象のテキストの分散表現の集合$X = \{x_i : \mathrm{Sim}_d(s_i) > n,\ i = 1, 2, \dots, N\}$に対して，k-means 法によるクラスタリングを行う。クラスタ数kは自動抽出する代表テキストの件数であり，パラメータとして事前に設定しておく。k-means 法における距離は，ユークリッド距離ではなく，コサイン距離を用いる。

提案手法では，集合Xから以下で定義する確率$p(x_i)$にしたがってランダムにクラスタ中心点の初期値を抽出する。ここで，$M_0 \subseteq X$は抽出済みのクラスタ中心点の初期値の集合である。$M_0 = \emptyset$ から開始し，$\#M_0 = k$となるまで繰り返すことで，クラスタ中心点の初期値を設定する。

$M_0 = \emptyset$ のとき：

$$p(x_i) \propto \mathrm{Sim}_d(s_i)^{\beta} \quad (x_i \in X)$$

$M_0 \neq \emptyset$ のとき：

$$p(x_i) \propto \left(\min_{x_j \in M_0} \mathrm{Cos}(x_i, x_j)\right)^{\alpha} \mathrm{Sim}_d(s_i)^{\beta} \quad (x_i \in X \setminus M_0)$$

$p(x_i)$は$\sum_{x_i \in X \setminus M_0} p(x_i) = 1$となるように正規化する。

$\alpha, \beta \geq 0$はパラメータであり，$\alpha = 2$，$\beta = 0$のときは通常の k-means++に対応する。$\beta > 0$のときは，類似テキストの件数$\mathrm{Sim}_d(s_i)$が大きいほど，ベクトルx_iがクラスタ中心点の初期値として抽出されやすくなり，s_iが代表テキストとして選ばれやすくなることが期待できる。

(5) 代表テキストの自動抽出

(4)の結果，クラスタC_hとその中心点$\mu_h (h = 1, 2, \ldots, k)$が得られる。代表テキストの自動抽出では，クラスタ$C_h$に分類されたベクトルのうち，コサイン距離で最も中心点μ_hに近いベクトルを選択し，そのベクトルを分散表現にもつテキストを代表テキストとして抽出する。以下，クラスタC_hの代表テキストを$r_h (h = 1, 2, \ldots, k)$とする。

(6) 最適なパラメータの選択

提案手法で用いる 5 種類のパラメータd, n, k, α, βの最適な値とその組み合わせをグリッドサーチにより探索する。パラメータを変えながら(3)から(5)の処理を繰り返し実行し，その結果をヒット率と AMI(Adjusted Mutual Information)[5] で評価する。ヒット率は以下で定義する。

$$\frac{1}{N} \#\{1 \leq i \leq N : \text{ある} r_h \text{に対して } \mathrm{Cos}(x_i, x(r_h)) \leq d\}$$

AMI は，クラスタリング結果の一致の度合いを表す指標で，0 から 1 の値を取り，1 に近いほど人手で作成したクラスタの正解データに近いことを表す。正解データは，分析対象のテキストの一部または全体を人手で分類することで，事前に作成しておく。ヒット率と AMI の合計値が最大となるパラメータの組み合わせを選択し，そのパラメータにおける代表テキストの抽出結果を出力する。

5. 評価

5.1 評価用データ

以下の 2 種類の評価用データを分析対象のテキストとして提案手法を評価した(Nは前処理後の件数)。クラスタの正解データには，分析対象のテキスト全体を用いた。

(1) データ 1(N=248 件)

・食品販売会社への顧客問合せを想定したデモデータ

・問合せ内容が端的に表現されている短い文

・前処理は，形態素解析による単語への分かち書きだけ

(2) データ 2(N=174 件)

・住宅メーカ A 社の社内システムの問合せデータ

・メールの文面のような長い文章で，定型的な文が多い

・前処理は，定型文・固有表現・不要語の除去，形態素解析による単語への分かち書き，抽出的テキスト要約による重要文 2 件の抽出

word2vec の学習コーパスには，2018 年 10 月 24 日時点の日本語 Wikipedia の記事全体のうち，50%の記事のテキストを利用した。語彙数は 299,961 単語であった。

5.2 評価結果

提案手法が 3.2 節で述べた 2 つの要件を満たすかどうかを確認するため，ヒット率と AMI を評価指標とした。パラメータの探索範囲とグリッドサーチで選択した値を表 1 に記す。また，データ 1 とデータ 2 で上記の探索範囲におけるパラメータの組み合わせごとにヒット率と AMI を算出してプロットした散布図を図 1 に示す。

表 1　パラメータの探索範囲と選択した値

	探索範囲	データ 1	データ 2
d	0.1, 0.15, 0.2, 0.25, 0.3	0.3	0.3
n	1 (※)	1	1
k	10, 25, 50, 75, 100	75	25
α	0.0, 1.0, 2.0	2.0	2.0
β	0.0, 1.0, 2.0	1.0	0.0

(※)クラスタの正解データを作成したときの値で固定

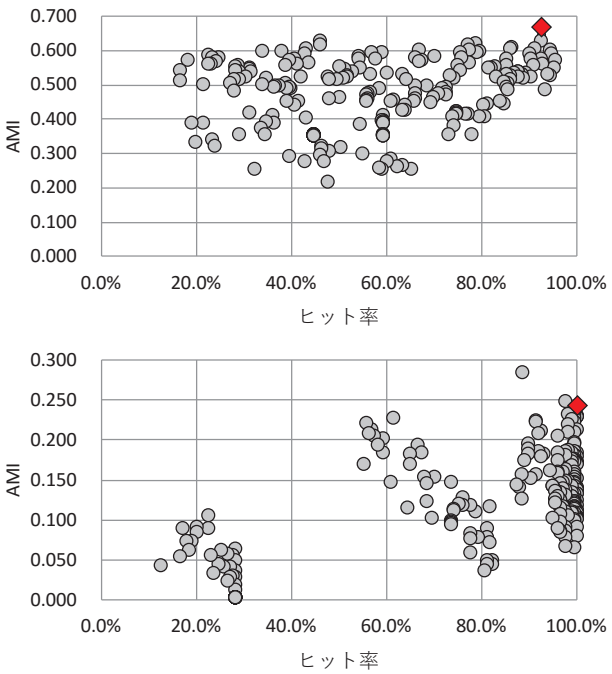

図 1　パラメータの組み合わせごとのヒット率と AMI(上：データ 1，下：データ 2)

パラメータの違いによってヒット率と AMI の値にバラツキがあることが分かる。高いヒット率を達成しつつ，人手で分類した結果に近いクラスタリング結果を得るには，適切なパラメータの選択が必要であることが分かる。

四角(◆)はヒット率と AMI の合計値が最大となるプロットを表す。このときのパラメータの組み合わせをグリッドサーチで選択して評価で用いた。選択したパラメータにおけるヒット率と AMI の値を表 2 に記す。

表 2　選択したパラメータにおけるヒット率と AMI

	データ 1	データ 2
ヒット率	92.3%	100%
AMI	0.649	0.237

ヒット率については，提案手法により 90%以上の高い値を実現できることを確認した。一方, AMI については，データ 2 で 0.237 と低い値となり，クラスタリング結果が人手で分類したときと近い結果とはならなかった。その原因として以下が挙げられる。

(1) 複数の話題を含むテキストがあり，抽出的テキスト要約では問合せ内容の特定が困難だった
(2) 複合語となっている業務用語が複数の単語に分かち書きされ，適切な分散表現を得られなかった

上記(1)に対応するためには，抽出的テキスト要約技術では不十分なため，テキストから問合せ内容を特定する文脈解析技術など，新たな技術を導入する必要がある。一方，(2)については，業務用語を形態素解析の辞書に登録した上で，その業務用語を含むテキストを学習コーパスに追加して単語の分散表現を学習するチューニングにより，解決可能と考える。

6. おわりに

本研究では，同義語の辞書が不要な FAQ 抽出技術として，テキストの分散表現を用いたクラスタリングと，クラスタから代表テキストを自動抽出する手法を考案した。本手法は，特願 2019-064867「文書分析装置および文書分析方法」として特許出願済みである。評価結果では，ヒット率は 90%以上となり，高いヒット率を実現できることが分かった。一方，AMI はデータによって低い値となり，人手に近いクラスタリング結果が得られない場合があることが分かった。改善には，テキストから問合せ内容を特定する文脈解析技術の導入や，業務用語を考慮したチューニングが必要と考える。今後は，複数の顧客と PoC を実施し，これらの改善策の効果を検証し，業務領域に応じた最適なチューニングを行えるようにノウハウを蓄積する所存である。

参考文献

1) (株)野村総合研究所，特許 5574842，ＦＡＱ候補抽出システムおよびＦＡＱ候補抽出プログラム
2) (株)日立ソリューションズ, 問合せ対応業務効率化・情報活用ソリューション (2019 年 8 月閲覧) https://www.hitachi.co.jp/products/it/magazine/hitac/document/2018/03/1803c.pdf
3) T. Mikolov, et al.: Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, pp.3111-3119 (2013)
4) 飯塚，他：コールセンターの音声認識テキストを用いたテキスト要約技術の研究，(株)日立ソリューションズ東日本 技報，第 24 号 (2018 年)
5) N. X. Vinh, et al.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, Journal of Machine Learning Research, Vol.11, pp.2837-2854 (2010)

飯塚　新司　　2008 年入社
ビジネスインキュベーション部
データ分析技術の研究開発
shinji.iizuka.zt@hitachi-solutions.com

菊地　大介　　2009 年入社
ビジネスインキュベーション部
テキスト分析技術の研究開発
daisuke.kikuchi.hz@hitachi-solutions.com

宮内　秀彰　　2008 年入社
Viz ソリューション部
CoreExplorer の提案・構築
hideaki.miyauchi.zd@hitachi-solutions.com

髙橋　毅　　2013 年入社
Viz ソリューション部
CoreExplorer の提案・構築
tsuyoshi.takahashi.rd@hitachi-solutions.com

黒澤　隆也　　2017 年入社
Viz ソリューション部
CoreExplorer の提案・構築
ryuya.kurosawa.ey@hitachi-solutions.com