問合せ・回答型テキストを対象とするテキスト情報検索の一方式

2V－4

森大二郎　杉崎正之　大久保雅且　田中一男
NTTヒューマンインタフェース研究所

1 はじめに

近年、多くの企業が顧客要求を吸収する情報源、あるいは顧客満足度を向上させるための手段として、ヘルプデスクと呼ばれる問合せ対応サービスの充実に力を入れている。ヘルプデスクにおける作業を効率化する手段の一つとして、ノウハウの流通と共有を促進するための技術が活発に研究・開発されている[1]。

ノウハウの流通と共有は、類型的な問合せと回答を抽出したQA集を作成したり、過去の全問合せ履歴をデータベース化して、応対要員から参照可能とすることによって実現される。履歴データベースは、QA集のように作成するための人的稼働を要さない反面、的確な検索指示を与えるのが難しいことが課題となっている[2]。

本稿では、電子化されたテキストとして問合せが与えられる場合に、問合せテキストそのものを入力として、履歴データベース中から内容が類似する問合せを検索するための一方式を提案する。

2 問合せテキストの特性

本稿では、企業等のヘルプデスクに対して不特定多数のユーザから電子メールやWWWの入力フォームを介して寄せられる、質問,要望,クレーム等の問合せと、これに対応する回答から構成されるテキストの集合を対象とする。

これと類似したテキストとして、質問と回答が頻繁にやりとりされるメーリングリストやUSENET上のニュースグループ(fj.questions.*等)の記事があり、これらを対象とした検索や分類を行う研究[3]がなされている。これらのテキストとヘルプデスクで扱われるテキストは、共に、既に電子化された状態で与えられ、問合せと回答の情報を多く含んでいる。また、内容が同様である問合せがしばしば繰り返される点も共通している。

しかし、同様の問合せが繰り返される現象はヘルプデスクの問合せテキストにより顕著に現れる。これは、ヘルプデスクにおけるメールが基本的に一対一の通信であり、情報が共有される契機がないためだと考えられる。同様に、文書のスタイルや語彙のばらつきについてもヘルプデスクの方がより顕著であるという傾向が見られる。このようなヘルプデスクの問合せテキストに特有の性質を以下に列挙する。

(1) 語彙や文書のスタイルがより多様である
同一の内容を表現するために、独自の語彙や言い回しが用いられることが多く、使用される語彙がまちまちである。挨拶やシグネチャの有無や順序、引用の形式等のスタイルもより多様である。

(2) 同様な問合せがより頻繁に繰り返される
内容が同一である問合せがより高い頻度で繰り返される。

(3) 回答テキストが安定して得られる
ほぼ全ての問合せに対して回答が返される。また、特定の対応要員から回答が返されるため、問合せと回答の判別が容易に行える。

問合せテキスト情報の検索では、テキストに含まれる単語や複合語から構成される特徴ベクトルのマッチングを用いる手法[4]が主に使用されているが、ヘルプデスクの問合せテキストにおいては主に(1)の特性が原因となり、検索精度を向上させることが困難となっている。

本稿では、このような特性を考慮し、問合せテキストそのものではなく、これに対応する回答テキストに着目することとした。

3 回答テキストを用いた特徴抽出

前章に述べたように、ヘルプデスクにおけるテキストにおいては、不特定多数のユーザから寄せられる問合せに対して、特定の少数のオペレータから回答が返される。したがって、表層上問合せテキストの語彙がまちまちであったとしても、内容が同一であれば、それに対応する回答テキストはより高い類似性を示すことが期待できる。この仮定に基づき、

An information retrieval method for texts consisting of pairs of inquiry and answer
Daijiro MORI, Masayuki SUGIZAKI, Masaaki OHKUBO, and Kazuo TANAKA
NTT Human Interface Laboratories

回答テキスト間の類似度によって、対応する問合せテキストの特徴ベクトルを調整することとした。
まず、テキスト t の特徴ベクトルを、

$$V(t)=(w_{t1},\cdots,w_{ti},\cdots,w_{tm})$$

$$w_{ti}=\frac{\log(\mathrm{TF}_{ti}+1)\cdot\log\frac{n}{\mathrm{DF}_i}}{\log length_t}$$

として算出する。ここで、n はテキストの総数、DF_i はテキスト集合における単語 i の出現回数、m はテキスト集合における全単語数、TF_{ti} はテキスト t における単語 i の出現回数、$lenghth_t$ はテキスト t の長さである。
文書集合は問合せと回答の対から構成される。問合せテキストの集合 Q の要素 q_i に対応する回答テキストの集合 A の要素を a_i とする。(図 1 参照)

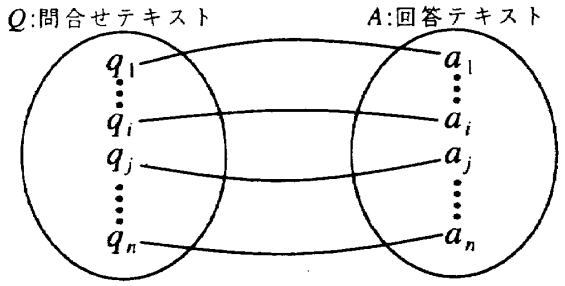

図 1: 問合せと回答の対応

回答テキスト間の類似度 R を以下のように定義する。

$$R_{ij}=\max\left(\frac{V(a_i)\cdot V(a_j)}{|V(a_i)|\cdot|V(a_j)|}-thresh,0\right)$$

$thresh$ は閾値として与える定数である。R の値に応じて対応する問合せテキストの特徴ベクトルを相互に加え、新たな特徴ベクトル V' を求める。

$$V'(q_i)=V(q_i)+\sum_{j=1}^{n}R_{ij}\cdot V(q_j)$$

検索時には、検索テキストに含まれる単語と各問合せテキストの特徴ベクトルとのマッチングを求め、点数の高いものを抽出する。

4 評価

ネットワークサービスにおける 3 週間分の問合せメールから 200 件の問合せ・回答テキストのセットを収集し、これに対して 30 件の問合せテキストをキーに検索を行なった。内容が同一である問合せ文を取得できた場合の適合率と再現率を求めた。回答テキストの類似性を加味した場合(V')としない場合(V)の再現率を 10%刻みにとり、それぞれに対応する適合率をプロットした結果を図 2 に示す。

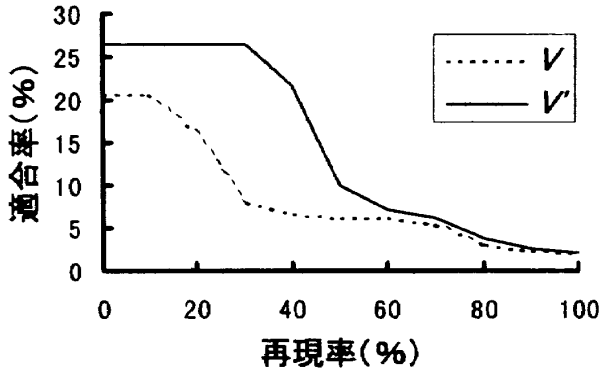

図 2: 検索の精度

回答テキストの類似性を加味した場合(V')の方が精度が向上していることが認められるが、問合せ内容が同一であるテキストを抽出するという条件が厳しいため、全体に高い精度が得られていない。精度をより向上させるためには、テキスト中の重要な単語に対して強い重み付けを付与する等の方法で、問合せの内容により促した特徴ベクトルを求める必要があると考えられる。

5 おわりに

問合せと回答の対から構成されるテキスト集合において、回答テキストの類似度に基づいて検索のための特徴ベクトルを算出する方式を提案した。今後はヘルプデスクにおける利便性を考慮し、検索結果として類似する問合せを列挙するのではなく、問合せ内容に基づいて分類した結果を提示する方法を検討する予定である。

参考文献

[1] Kondo: Business Process Management in Customer Contact Services, Network Operations and Management Symposium,1998
[2] 多田和市: 情報を効率よく引き出す検索ソフト, 日経ビジネス,1998 年 6 月 29 日号
[3] Sato: Natural Language Processing in the Automated Editing System for the USENET Articles, Proceedings of the Natural Language Processing Pacific Rim Symposium,1997
[4] Salton: Automatic Text Processing, Addison Wesley, 1989