問合せ回答型文書集合における類似事例抽出方法の一検討

3U－5

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、企業において顧客に対する問合せ対応サービスを充実させようとする動きが盛んになっている。多くの企業では、ヘルプデスクと呼ばれる顧客対応部門を設け、作業やノウハウの集約と合理化によって業務を効率化すると共に、サービスの品質を向上させ、顧客の信頼と満足度を得ようとしている。また、これらのサービスを通して得られる情報を分析し、顧客要求を推定するための情報源として活用する試みも盛んになされている。

本稿では、ヘルプデスクにおけるこれらのデータ処理への適用を想定し、問合わせテキストからその内容に即した特徴量を算出して類似事例や関連語句を抽出する手法について報告する。

2 問合せ文書の特徴

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。

一般に、ヘルプデスクに寄せられる問合せテキストは不特定多数のユーザによって書かれるため、同一の概念が様々な語彙で表現され、テキスト集合全体としては同義語や多義語が顕著に現れるという特徴を有している。

テキスト情報の特徴を表現する手段としては、テキストに含まれる単語を軸とするベクトル空間モデルが主に用いられるが、テキスト集合に多くの同義語や多義語が含まれる場合には、特徴ベクトルがテキストの内容を的確に反映しなくなるという問題点がある。

筆者らは、この問題に対処するために、問合せテキストではなく、これに対応する回答テキストに着目して文書の特徴量を計算するというアプローチを取っている。

ヘルプデスクにおける問合せ回答テキスト集合においては、不特定多数のユーザが記述する問合せテキストに対して、特定少数のオペレータが回答テキストを記述するため、両者とも同一の問合せ内容に言及しながら、使用される語彙の分布には非対称性が見られる。すなわち、回答テキストの方が、同一の内容に対しては、より一様な表現を用いるという傾向が見られる[1]。従って、回答テキストに含まれる語彙に基づいてテキストの特徴量を求めることによって、問合せ内容の類似関係をより的確に反映した特徴量を得ることができる。

また本手法においては、関連性の高い文書および単語をベクトル空間の近傍に配置する、潜在意味分析(Latent Semantic Analysis)[2]手法を適用した。

潜在意味分析では、単語×文書からなる行列に特異値分解を施すことによって、共起性の高い行および列を縮退し、次元数の少ないベクトル空間上への写像を求める。これによって、関連性の高い文書および単語を共通のベクトル空間の近傍に布置することができる。このため、同義語と多義語の問題を一部解消すると共に、単語の特徴量と文書の特徴量とを相互に演算することが可能となる。

3 問合せ・回答型文書に即した特徴量計算手法

本手法では、問合せテキストと回答テキストとが一対一に対応するテキスト集合が予め相当数蓄積されていることを前提としている。

特徴量計算手法の骨子は以下のとおりである。

1) 回答テキスト集合における単語×文書行列に潜在意味分析を施し、問合せ内容に即した特徴ベクトルを得る。
2) 各回答テキストに対応する問合せテキストに含まれる語彙を同じベクトル空間に配置し、その重心を求めることによって各語彙の特徴ベクトルを算出する。

次に本手法の詳細を解説する。

まず、各回答テキストを形態素解析した結果に基づいて、$t \times d$の行列 $A=(a_{ij})$を求める。ただし、

A study on extracting similar cases from inquiry and answer text pairs.
Daijiro MORI, Masayuki SUGIZAKI, Masaaki OHKUBO, and Kazuo TANAKA
NTT Human Interface Laboratories

$$a_{ij} = \frac{\log\left(\mathrm{TF}_{ij}+1\right)\cdot\log\dfrac{d}{\mathrm{DF}_i}}{\log length_j}$$

とする。ここで、t は回答テキスト集合における全単語数、d はテキストの総数、TFij は回答テキスト j における単語 i の出現回数、DFi は回答テキスト集合における単語 i の出現回数、$length\ j$ は回答テキスト j の長さである。

行列 A の特異値のうち、大きいものから順に m 個の要素を対角成分とする対角行列を S、これに対応する右特異ベクトル行列を D とする。

このとき、テキスト j の特徴ベクトル V_j を以下によって与えるものとする。

$$V_j = D_j S$$

ただし、D_j は、行列 D の j 行目の行ベクトルである。

次に、問合せテキストについても形態素解析を行い、$u \times d$ の行列 $Q=(q_{ij})$ を求める。ただし、

$$q_{ij} = \frac{\mathrm{TF}_{ij}}{\mathrm{DF}_i}$$

ここで、TFij は問合せテキスト j における単語 i の出現回数、DFi は問合せテキスト集合における単語 i の出現回数である。u は、問合せテキスト集合における全単語数である。

新たな問合せテキスト x の特徴ベクトル Vx は以下の行列の積算により求められる行ベクトルとして算出する。

$$Vx = X^{'}QDS$$

ここで、$X=(xi)$ は u 次元数ベクトルであり、

$$x_i = \frac{\mathrm{TF}_i}{length}$$

とする。ただし $length$ は問合せテキスト x の長さ、TFi は問合せテキスト x における単語 i の出現回数である。

問合せテキスト x と類似するテキストの検索は、$\left|V_x - V_j\right|$ を最小とするテキスト j を探索することによって実現する。

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。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 1 に示す。

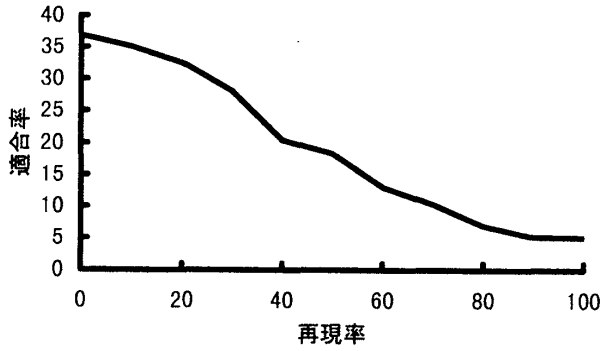

図 1 検索精度

本実験データにおいては、問合せ内容が同一であっても、問合せテキストに共通の単語が含まれない場合も存在し、内容が同一である問合せ文を取得することは、ごく厳しい条件であったと言える。本手法は、このような実践的なデータセットにおいて 1/3 程度の適合率を実現しており、ヘルプデスクにおける回答支援等に適用可能なレベルに達していると考えられる。

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を反映した特徴ベクトルを構成し、これに基づいて問合せテキスト及びこれに含まれる語彙の特徴ベクトルを算出する手法を提案した。今後は、類似文書の抽出精度の更なる向上に取り組むと共に、テキストの分類、関連語/同義語の抽出等への適用性を検証する予定である。

参考文献

[1] 森大二郎, 杉崎正之, 大久保雅且, 田中一男: 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, 情報処理学会 第 57 回全国大会. 2V-4, 1998

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman: Indexing by Latent Semantic Analysis, journal of the american society for information science. 41(6):391-407, 1990