モダリティ分析に基づく自由回答アンケートの分類

乾 裕子　　内元 清貴　　井佐原 均
計量計画研究所　　郵政省通信総合研究所
hinui@ibs.or.jp　　{ uchimoto, isahara }@crl.go.jp

1. はじめに

近年、市民参加・合意形成などの社会潮流を反映し、自由回答によるアンケート調査が注目されている。自由回答は、調査主体のコントロールが少なく、率直な世論を得ることができる。また、ホームページを利用してインターネット利用者に意見要望を自由に書き込んでもらう試みも行われている[5]。

一方、これらの調査では回答数が膨大であり、得られた回答に対する分析の労力が多大であるため不十分な分析に終わる場合、あるいは収集した回答が無駄になる場合もある。自由回答の意見集約は手作業によるものが多く、自動分類により人的コストを軽減できる可能性が高い。

また、人の判断による分類は客観的基準になりにくいという問題がある。調査結果や、その分析結果を一回限りのものでなく継続的に蓄積するためには、できるだけ作業者に依存しない基準を作ることが重要と考える。

以上の背景を踏まえ、本調査研究では工学分野で研究開発されているテキストクラスタリング技術を用いて大量の自由回答を適切に自動分類するための手法を検討する。同時に、言語学的分析を行い、アンケート調査における自由回答を対象に人が自由記述で意思表明する際の表現型に関する網羅的分析、および意図抽出のモデル化を行う。将来的には、意図抽出モデルによる自由回答テキストクラスタリングシステムの構築を目指す(図1)。

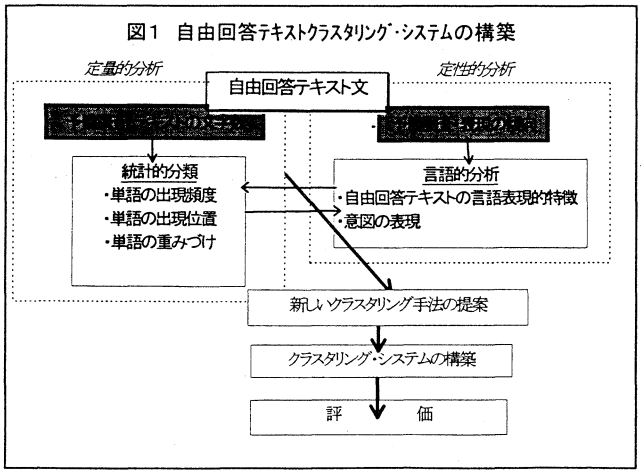

図1 自由回答テキストクラスタリング・システムの構築

本稿では、図1で網掛けされた予備調査についてその結果を報告する。

2. 関連研究との比較

アンケート調査に関する研究は、社会科学系の研究分野、および心理学系の研究分野を中心に行われている。しかし、自由回答は調査主体がコントロールしにくく得られた結果をまとめにくいことから分析対象となることが少なかった。このような状況の中、自由回答に対する関心の高まりから自動分類を目指す研究が現れ始めている[3][5][6][7]。しかし、本研究で回答者の意図の分類を検討するのに対し、これらはキーワードに着目した従来の内容分類に近い。また、多分野に渡って行われている意識調査分析も、キーワードに着目した従来型のものである[8]。近年の合意形成に見られる社会潮流を反映したという点では、複数の参加者による議論を集約する支援システムの研究開発がある[4]。意見の選択過程において意志決定者が主観に基づき意見を選択していく点に、客観的基準を作成する我々の立場との相違がある。

新聞の自動分類と異なり、アンケートの回答では回答者の意図を把握するために分類を行うことが大きな目的である。したがって、名詞を中心とするキーワードを手がかりにした分類は、自由回答の分類には適当でない。以上の考えは、実験的調査を経て得たものである。実験的調査については4節で述べる。

3. 調査対象

ここでは、本研究で対象とした自由回答テキストについて説明する。われわれが使用したキックオフレポートとは、道路審議会基本政策部会の「21世紀の道を考える委員会」が平成8年5月から7月末に実施した全国規模のアンケート調査である。将来的な道路計画に市民の声を活かす目的で行われている。キックオフレポートの回答人数は 35,674 人、回答数(意見数)は 113,316 件である。

意見は、ハガキ、封書、FAX、電子メールによる回答の他、ホームページへの書き込みによって集められている。回答方

法を下記に示す。

・あらかじめ設定された道づくりに関する12のテーマの中から、回答者は関心の高いテーマを選択する
・各テーマに対し、4個程度の参考意見、およびグラフや図などの参考資料が提示されている。
・120 字程度の文字を記入できる回答欄に意見を書く。意見が多いときは別紙に記入する。

設定されたテーマは次のとおりである。

1)くらしと道のかかわり 2)渋滞の解消 3)市街地と道づくり 4)生活環境と利便性 5)交通安全の確保 6)道づくりと合意形成 7)情報通信技術と交通 8)国土の使い方と機能配置 9)地域づくりと生活サービス 10)高速道路の料金と道路整備 11)道路空間の使い方 12)民間と行政の役割分担

以上の方法で集められたキックオフレポートの回答はボイスレポートと呼ばれる。集められた回答は、自由回答の従来の集計同様、人手作業により12のテーマに分類されている。分類は基本的に回答者の選んだテーマ番号で行われるが、テーマ選択が誤っているとみなされる回答に対しては、分類番号にテーマ番号と異なる番号が付けられている。分類番号とテーマ番号が異なるのは全体の約2%である。回答者および分類作業者が12のテーマに選別できない回答は10%程度ある。

4. tfidf 法による実験的調査

ボイスレポートを概観するために、1000 字以上 1500 字以下のテキストに対して、名詞に着目した tfidf 法によるクラスタリングを行った。テキストのクラスタリングには融合法の一つであるグループ平均法を用いた。これは最も単純なクラスタリング手法の一つである。テキストの内容は、母集団に出現する全名詞をベクトルの要素としてベクトル空間モデルで表し、各テキストの各ベクトル要素にはtfidf法に基づいて得点を与える。特定のテキストに頻出するが、母集団全体では頻度が低いような語はテキスト分類に有効であると考えられる。tfidf 法を用いると、このような語に高得点を与えることができる。

クラスタリングでは、最初、一つ一つのテキストをそれぞれクラスタと考え、それぞれのクラスタ間の距離はベクトルの cos 値を計算することにより、1-cos として与える。ここで、距離(1-cos)が小さいものほどクラスタの内容が近いと考える。その後、距離が最小の二つのクラスタを一つのクラスタに融合するという操作を繰り返す。一つに融合されたクラスタと他のクラスタとの距離は、グループ平均を求めることによって与える。最終的に、クラスタ間の距離の最小値がある閾値を上回る場合にクラスタの融合を止める。

最終的に残ったクラスタのそれぞれに含まれるテキストは、母集団に一般的でない名詞のうち、多くを共通に持つと期待できる。このように名詞に着目してクラスタリングした結果が自由回答アンケートの分類として適切かどうかを調べた。

単語の出現頻度による分類に関して問題が生じるのは表1の例である。

①同じ分析番号の中に賛成反対の例がある場合

表1 単語の頻度計算による結果と目指す分類結果

	1	2	3	4		
	回答例	ﾎﾞｲｽﾚﾎﾟｰﾄの分析番号	単語の頻度計算	提案手法による結果(案)		
				内容	賛否	提案・要求
①	高速道路の料金が、値上がりするのは仕方のないことだ。それよりもっと、もっと高速道路をふやし、便利がよくなるようにした方がいいと思う。	10:高速道路の料金と道路整備	0.3914	高速道路		料金値上がり支持
	道路料金に依って便利な道路ができている。料金が高いということは当を得ない。	10:高速道路の料金と道路整備				料金値上がり不支持
②	環境破壊が進んでいるので緑を増したほうがいいと思います。	11:道路空間の使い方	0.3298	環境破壊		提案:緑を増やす
	環境破壊はよくないが、道路を作るためには仕方のないことだと思います。	1:くらしと道のかかわり				道路作り支持
③	Cさん、Dさんと同様で、電柱等を地下へ納めると歩道も広くなるし、景観も良くなると思う。上空にいろいろと建設して空が見えなくなるのは嫌なので、地下部をもっと利用して発展させて欲しい。	11:道路空間の使い方	0.6355	道路空間	賛成C,D	地下利用
	大通から札幌駅までの地下歩道建設が無理であれば、空中歩道の建設をお願いしたい。	11:道路空間の使い方				空中歩道

②異なる分析番号だが、同じテーマに対して異なる意見を述べている場合

③同じ分析番号の中に異なる提案がある場合

アンケート調査の回答分類では、回答の賛否を把握することがもっとも重要である。したがって、賛否の表明は分類基準とする。また、②については、提案する手法の中で名詞の頻度計算を併用することにより、同じテーマに対する見解の異なりであることが導けると考える。賛否の把握と同様に重要である提案・要望を分類基準とする(③)。

この実験的調査を通じ、人手作業あるいは名詞の頻度分類による問題点を指摘することができた。

5. モダリティ分析

回答者の意図を把握するために、本研究は従来のテキストクラスタリングでは取り上げられることのなかったモダリティに着目する。モダリティは、「現実との関わりにおける、発話時の話し手の立場からした、言表事態に対する把握のしかた、及びそれらについての話し手の発話・伝達的態度のあり方の表し分けに関わる文法的表現」と定義され[2]、書き手の意図が文法的に現れたものと考えられている。したがって、モダリティ表現に着目して自由回答を分類することは適切であると考える。

また、一般に助動詞及び助動詞に相当する語がその機能を担うと考えられており、実際の用例から助詞・助動詞相当の表現を集めた森田・松木の研究では 180 あまりの助動詞相当の表現を、禁止・義務・可能・許可・推量・提案・要求など 14 の意味に分類している[1]。この分類例を参考に、実際の自由回答ではモダリティ表現がどのように現れるか調査する。

5. 1 調査方法

あらたにモダリティに着目して分析を行うため、調査方法の見直しをはかった。モダリティ表現を中心に自由回答の特徴を知るため、その分析対象を以下の方法で抽出する。

①テキストの文字数

テキストの特徴を把握するために、113,316 件の全文字数を計算したところ、100 字未満の回答が全体の 66%を占めていた(図2)。400 字未満までの回答でほぼ100%近くに達するが、それより長い回答では、400字以上1000字未満が1165件、1000字以上3000字未満が 182 件、3000 字を越える長い回答も 27 件あった。

上記の結果から、100 字～200 字程度の回答がもっとも多いことが判明した。これは、200 字未満のテキストの特徴を把握することの必要性を示す。

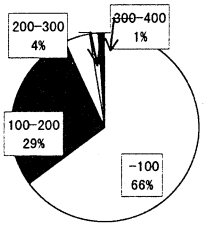

図2 回答テキストの特徴
(N1-N2:N1 字以上 N2 字未満)

②表現の傾向

全回答数 113,316 件の母集団から 1,000 件の標本データを層別サンプリングによってランダムに抽出した(現時点では、そのうち 100 字未満のテキスト 667 件について調査を進めている)。

100 字前後の回答を概観すると、下記の表現が多く見られる。

・賛成および反対の表明(…に賛成、…と同じ意見、…の意見の通り、…に共感する etc.)

・提案あるいは要望表現(…してほしい、…もらいたい、…した方がいい、…すべきだ、…が重要、…が大事だと思う etc.)

③先行研究との比較

対象とした 667 件の回答に、先に挙げた松木・森田の 180 あまりの助動詞相当表現が現れているかどうか調べる。

5. 2 調査結果

アンケート調査の自由回答テキストでは、従来注目されていなかったモダリティ表現に着目することが重要である。回答者の意志が現れた表現型に着目することで、テキストクラスタリングを行う際、より高い精度を得られる可能性がある。

667 件のデータに現れていた表現形のうち 269 件が、森田・松木[1]の挙げた例にマッチしている。表現形に付けられたモダリティの意味(＝心的態度)は、依頼、意志、勧告、義務、許容、限定、提案、程度、当然、要求、推量である。しかし、実際の回答例を観察すると、必ずしも分類された意味の違いが見られず、表現に意味を固定的に対応させるのは言語処理のうえで好ましくない。これは、国語学の従来研究では対象分野を決めることなく一般性を求めていたためと考えられる。

提示された表現形では実際の例を網羅できず、従来型の助詞・助動詞相当語を基にさらに 100 個の表

表2 従来型のモダリティと広義のモダリティの対照

モダリティ		回答	広義のモダリティ	
心的態度	表現形		心的態度(意図)	表現形
意志	ようにする	人や車の安全のためには、広い道を作り、高速道路のように、人が出入り出来ないようにすることが一番だと思う。	提案・要望	出来ないようにすることが一番だと思う
要求	てほしい	クルマの安全性の向上は大事だと思う。それだけでもずいぶん安全になるだろう。歩道がせまいのでもう少し広げてほしいな。(そこに自転車もくるから)		は大事だと思うてほしい
限定	外にない	幹線道路で歩道のない道路はいっぱいあるが、ネックになっているのは用地取得の困難にある。問題は住宅地の中に工場等が非常に多い。工業地域と住宅地域を短期長期に分けて町づくりをする外にない。		ほかにない
勧告	たいものだ	道幅(車道)を広くし、歩道、自転車用に道を使い分けるようにしたら事故が少なくなるのではないか。一般住宅内の道幅も大変狭く感じる。冬の季節を考慮した道路を行政、民間で考えたいものです。		のではないか考えたいものだ
許容	てもよい	(民間と行政の役割分担について)市街地空間を有効利用できるような民間の道路整備があってもよい。		があってもよい
要求	てもらいたい	車中心の社会とは言っても、老人、子供や身障者にも優しい道をつくってもらいたい。弱い立場の人々に優しい道路が必要だと思う。		もらいたい必要だと思う

現を加えた。また、従来、モダリティとはみなされないが、アンケート調査に特徴的と考えられるいくつかの表現に着目し報告する。

表2は、アンケート調査の自由回答という場面においては従来型モダリティの範囲外である「大事だと思う」「必要だと思う」「のではないか」といった表現が「べきだ」「てもらいたい」等の表現と同じ働きをすることを示す。われわれは、モダリティの意味と表現形について言語処理の目的に合わせて変化する動的な枠組みで捉え、これを広義のモダリティと考える。広義のモダリティには、先に挙げた動詞句や形容詞句も含める。667 件のデータには、本研究で広義のモダリティとみなした表現が 87%現れている。アンケートの自由回答に特徴的なこれらの表現についてさらに詳細に調べるため、下記を仮説として分析をすすめる予定である。

仮説 1.分析対象がアンケートの自由回答という限られた範囲である場合、一般に定義されているモダリティ表現や動詞よりも意味が限定される。

仮説 2.形容詞や動詞にも助動詞・助動詞相当語同様、いわゆるモダリティの機能を担う表現がある。

6. おわりに

自由回答の自動分類においては、従来型の名詞を中心としたキーワード分類だけでなく、モダリティに着目することが重要である。モダリティは、回答者の意図が文法的機能として表現に現れたものである。したがって、回答の意図を把握することが重要な自由回答の分析には効果的であると考える。

今後は、さらにボイスレポートに基づきモダリティの詳細な分析をすすめ、名詞を中心としたキーワード分類との統合、システム構築について検討する予定である。

謝辞: 本研究を始めるにあたり、ご助言いただきました敬愛大学の高橋和子先生に感謝いたします。

参考文献:

[1]森田・松木:『日本語表現文型』アルク, 1989
[2]仁田義男:『日本語のモダリティと人称』ひつじ書房, 1991
[3]野崎ほか:アンケートにおける日本語自由文の情報分析, 情報処理学会第 47 回全国大会論文集, 3,pp165-166, 1993
[4]Nikos Karacapilidis, et al. "Collaborative environmental plannning with GeoMed", European Jounal of Operational Reseach 102,pp335-346.1997
[5]醍醐朝美:電子調査法による自由回答形式の意見聴取データの解析,日本行動計量学会第24回大会発表論文抄録集, pp58 –59,1997
[6]高橋和子:自然言語処理によるSSM職業コーディング・システムについて,日本行動計量学会第24回大会発表論文抄録集, pp166-167,1997
[7]大隅ほか:自由回答データの解析法についての提案,日本行動計量学会第24回大会発表論文抄録集,pp176-179,1997
[8]須賀・大井:自由回答記述データを用いた瀬戸大橋に対する住民意識の解析, 土木計画学研究・講演集 No.20(2), PP31-34,1997