疑問型表現自由回答データを用いた 社会資本整備に対する市民の関心の抽出方法に関する基礎的研究*

Fundamental Study of Abstracting Citizens' Interest toward Infrastructure Project by Using Interrogative Sentences of the Open-Ended Text Data*

福田大輔**・庭田美穂***・屋井鉄雄****
By Daisuke FUKUDA**, Miho NIWATA*** and Tetsuo YAI****

１．はじめに

(1) 研究の背景

社会資本整備事業の計画段階から市民の声を聴き，計画策定に反映させることを目指した“パブリックインボルブメント(PI)”の代表的な意見収集手法として，アンケート調査，FAX，電話，はがき，インターネット等が挙げられる．上記のような従来的なものに加えて，行政が住民との対話をしながら意見を引き出すオープンハウス，公聴会，ワークショップなどの場面における意見収集など，近年，その方法は多様化の様相を帯びている．これに伴い，市民の意見を直接表明した「自由記述アンケート」から得られた市民意見を用いて，道路整備に対する市民側のニーズや不満度を抽出する研究[1),2)]も行われている．

PIにおいては，いかなる意見収集手法を用いても，得られた自由意見は最終的には「自由回答テキスト」の形となり，市民へのとりまとめ結果の公表や計画プロセスにおける参考意見として重要なデータとなる．しかし，それらの自由回答テキストを集約した後に，どのようにまとめるのかについての確立された方法や基準はない．実際のPIの現場では，市民からの意見が極めて膨大な量のテキストデータとして残され，行政官やコンサルタントが1つ1つの意見をあるキーワード(例えば“道路”，“環境”など)に基づき，人手で分類しているのが現状である．そのため，①時間とコストが費やされる，②自由回答を分類する個人の主観の影響を受ける，③結果，本来住民が言わんとしていることや，テキストに含まれた回答者の意図や関心を無視してしまうことなどが課題として残されている．例えば，本研究で対象とする“(仮称)横浜環状北西線”構想段階におけるPI導入事例でも，2005年3月時点までに4587件という大量な意見が集められ，それを人手で分類してきた．そのため，記述内容別の分類程度に留まっており，テキスト表現に含まれた市民の意図や関心に注目した分類はなされていない．

矢嶋[3)]が指摘するように，PIにおいては，多数の「意見を集める」だけでなく，市民等の利害関係者が表明する意見あるいは示す態度の要因となっている懸念や不安等，深い「関心」の所在を明確にし，それを行政だけでなく市民全体が認識することが重要である．そのためには，市民側から出される大量な自由意見をPIプロセスの中で埋没させないように配慮すると同時に，各意見に秘められた人々の“関心”が何であるのかを，効率的に抽出・分類することを可能にするような手法を確立させることが重要である．このような手法の確立により，将来的には，行政の意見集約作業の支援ツールとして資することが期待される．

(2) 研究の目的

以上のような問題意識のもと，本研究では，次の三点を目的として分析を行う：①自然言語処理技術を援用して，自由回答テキストに含まれる市民の関心を軸にしたテキストの自動分類を行う．②関心の相違が顕著である「疑問型表現文」に着目し，分類結果から，疑問型表現文による意見表明における市民の関心の所在を明らかにする．③意見を受取る側の影響を知るため，読み手による関心の解釈の傾向を明らかにする．

①に関しては，機械学習（Support Vector Machine）に基づく既存のテキストマイニング手法を援用し，PIにおける自由記述データの自動分類の可能性について実証的に検討する．②に関しては，既往研究と異なり，PIの場面においてより慎重に取り扱うことが必要とされる「疑問型表現文」について詳細に分析する．③に関しては，「疑問型表現文」の解釈の多義性について，実験データを用いて詳細に検討する．以上，①～③の検討を行う点が，本研究の特徴である．

*キーワーズ: 自由回答テキスト, PI, 自動テキスト分類
**正会員　博(工)　東京工業大学大学院理工学研究科
（〒152-8552 目黒区大岡山 2-12-1-M1-11　Tel: 03-5734-2577）
***非会員　修(工)　株式会社インテリジェンス
****正会員　工博　東京工業大学大学院総合理工学研究科

(3) 既往研究の整理

PI における自由記述データを対象として筆者らも研究を行っているが[1),2)]，これらの研究では，プリコードデータと自由記述データの対比及び連関分析に留まっている．また，計画策定プロセスにおける自由記述データを分析したものとして，自治体総合計画の計画過程において一般市民と専門家の発言にはどのような違いがあるのかを把握したもの[4)]，本研究同様，(仮称)横浜環状北西線計画の構想段階で得られた自由回答アンケートを用いて回答属性(フェイス項目)と自由回答との関連性を見たもの[5)]等が見られるが，これらは，大量な自由記述データの効率的な分類を意図していない．本研究の立場に近い研究として，道路計画についての自由記述回答から“要求意図”を取り出し，その情報をもとにテキストとして回答を自動分類した大塚[6)]が挙げられるが，使用されている自由回答データは，はがき，FAX，電子メールから得られたもののみであり，対話によるコミュニケーション(説明会等）から得られた文章は対象としていない．

これらの既往研究に対し，本研究は，プロジェクトの構想段階 PI において，多様なメディアから得られた自由記述回答において多く見られる“疑問型表現を文末表現に含むテキストデータ”に対象を絞って分析を行う．後述するように，一口に「疑問型表現文」と言っても，内に秘められた市民の「関心の所在」は多様であり，これを効率的に分類することは，PI の実務においても意義のあることと考えている．

２．対象とする事業計画の概要

(1)（仮称）横浜環状北西線計画の概要

（仮称）横浜環状北西線（以下，北西線）は，横浜市域の幹線道路での渋滞緩和へ向け，第三京浜道路（港北 IC）と東名高速道路（横浜青葉 IC）間を結ぶ区間に計画であり，2003 年 6 月 26 日より検討が開始されている（**図-1**）．構想段階から PI を導入して計画を進行した事例として知られており，具体的な意見徴集手法として，オープンハウス，「地域住民の意見を聴く会」，「周辺自治会・町内会との会合」，E メールや web（ホームページ），アンケート調査，はがき，フリーダイヤル，FAX が用いられている．また，担当行政機関への来所による意見受付も行っている．構想段階は「北

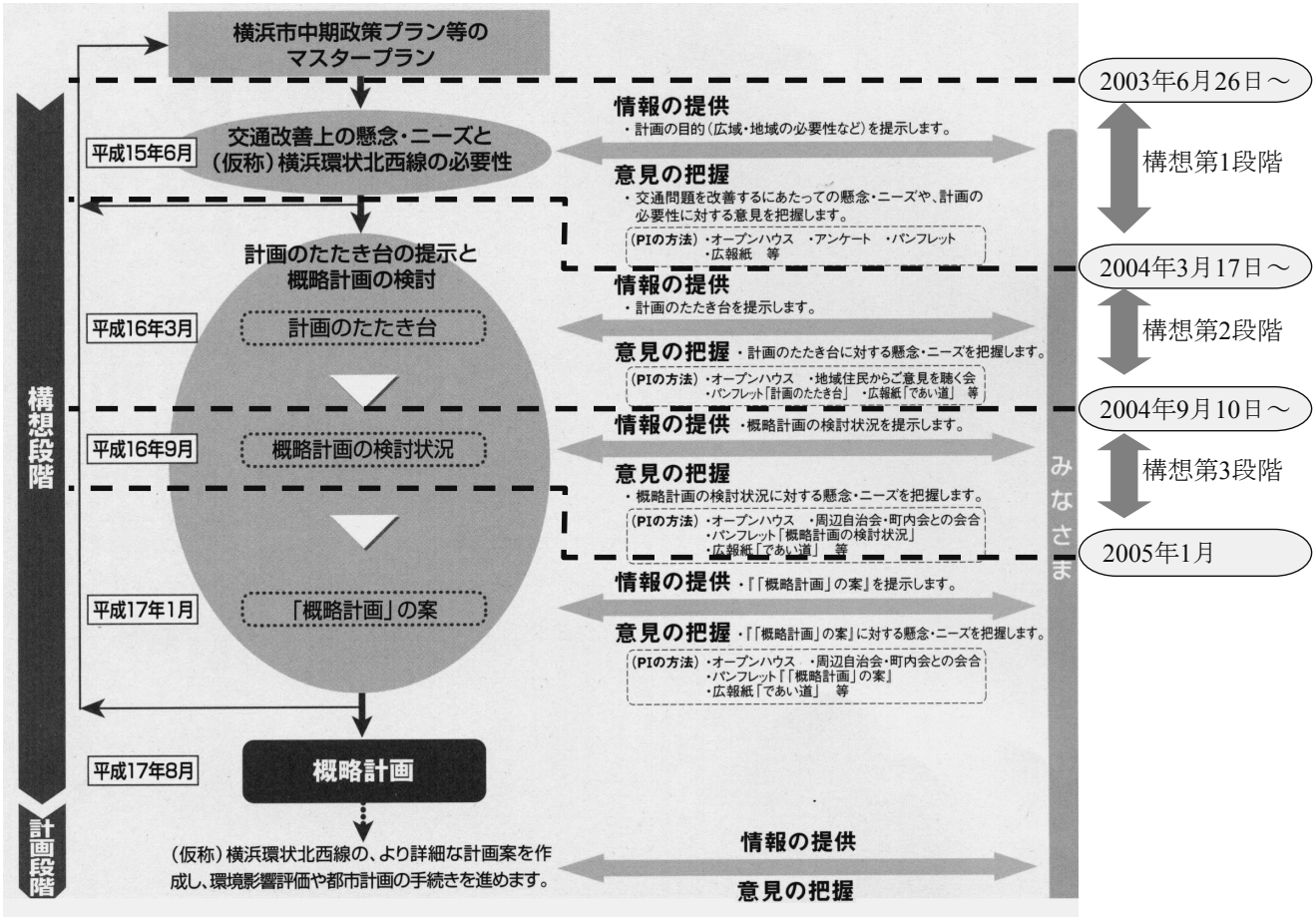

図-1 (仮称)横浜環状北西線－構想段階 PI プロセス（北西線 HP より引用，一部加筆）

西線の必要性や懸念」を問う第 1 段階から,「計画のたたき台と概略検討」を問う第 2 段階を経て,「概略検討の状況」をまとめる第 3 段階に分けられる.

⑵ 用いるデータ

本研究では, 上記の各手法を用いて収集された①構想第 1 段階(2003 年 6 月 26 日～2004 年 3 月 16 日)と②第 2 段階(2004 年 3 月 17 日～2004 年 9 月 9 日)において得られた自由回答テキストのうち，後述する「疑問型表現文」を含む意見を分析対象データとする．それぞれの期間における疑問型表現文の数は，①が 934 文，②が 581 文で，合計 1515 文となっている．

３．使用する語句の定義

⑴ 疑問型表現文の定義

本研究では, 言語論におけるモダリティの概念に基づいて疑問型表現文を定義する. そのため, まず,「疑問文のモダリティ」に関して概説する.

文の叙法性＝モダリティ (modality) の概念は，英語の研究では must，may，can などのいわゆる法助動詞 (modal auxiliary) の表す意味をモダリティとするのが一般的である．宮崎ら [7] によると，モダリティは**図-2** に示すような体系で表されることが一般的である．一方，日本語文の基本構成は，客観的な事柄を示す“言表事態”と，主観的な判断や態度を表す“言表態度”に分かれるとされている．その例を**図-3** に示す．仁田・益岡 [8] は，この言表態度を“モダリティ”と定義している.

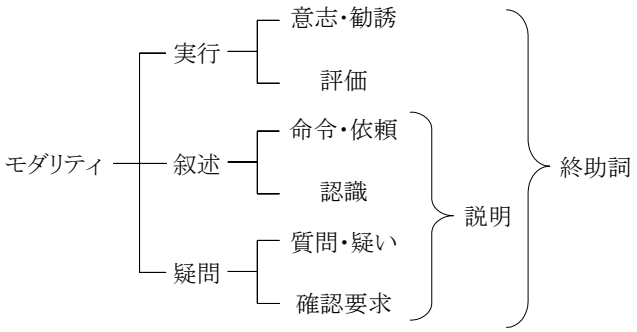

図-2　モダリティの体系 [7]

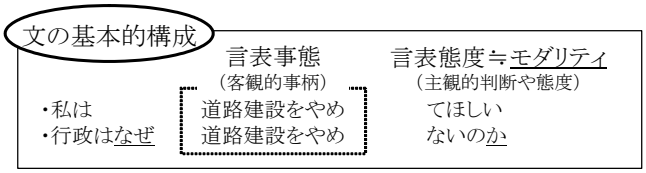

図-3　日本語文章の基本構成

本研究では，仁田・益岡 [8] によるモダリティの定義を採用する．その上で, 分析に用いる「疑問型表現文」を,【言語論における『疑問文』, すなわち，疑問型のモダリティを持つ文に，その省略形や方言形，文末に疑問符がついた文を含めたもの】と定義する．ここで，“疑問のモダリティ”とは,

・「疑問代名詞＝だれ・いつ・どこ」,
・「疑問数詞＝いくつ・いくら」,
・「疑問副詞＝なぜ・どう」,
・「疑問連体詞＝どの・どんな」,
・「～のではないか(否定疑問)」,
・「～ではないか，～だろう(確認要求)」,
・及びそれらの丁寧形；

が含まれる文である．本研究ではさらに,

・上記の定義を基本とした省略形や方言形
（例. ～では？・～ちゃいます？),
・末尾に助動詞「～か」が表れるもの
（例. ～するつもりか),
・「～ね. 」「じゃない？」「～でしたっけ？」
といった口語体,
・その後に同意や確認を求めるような表現
（例. ～と思いますが.),
・疑問副詞に直接疑問符がついた表現
（例.「何？」「いつ頃？」）

を含む文に関しても，疑問型表現文と考える．

なお, 本研究で分析対象とする北西線構想段階における対話型コミュニケーション手法の 1 つである『地域住民の意見を聴く会』の発言内容を精査すると，1 つの発言（文）に対し，話し手（住民）と受け手（行政）の間で理解の相違が見られる場面が散見された. 特に『疑問型表現』の発言においてそれが顕著であり，住民が“意見”として表明したものを, 行政は“質問”と捉えてしまい, 無理に回答を行うという状況が多数見られた．このように，PI における疑問型表現自由回答データは, 回答者が様々な“含み”を持たせた意見表明の形になり易く, その類型化を的確に行う必要性が高い．このような理由から，本研究では疑問型表現を文末表現に含む自由回答に, 特に分析対象を絞っている.

⑵ 関心の定義

“関心”とは,「物事に興味をもったり，注意を払ったり，気にかけること [9]」と定義される．また，PI においては，交渉学（例えば，Fisher & Urly[10]）における定義に基づき,「利害関係者が表明する意見ある

いは示す態度の要因となっている“懸念や不安等”」を“関心”と見なすことが一般的である[3].

以上に配慮し, 本研究で着目する疑問型表現文においては, 以下に示される 8 つの分類軸に基づいて, 市民の“関心”の所在が表されるものと仮定する：

≪質問≫・≪疑い≫・≪確認≫・≪要求≫・
≪不満≫・≪懸念≫・≪賛成≫・≪反対≫

各分類軸の定義を**表-1** に示す. 言語論における「疑問文」は, ≪質問≫や≪疑い≫を表すことが主な機能とされているが, 本研究では, たとえ同じ表現型であっても, 他の様々な“関心”を表している可能性があると考え, 上記の候補を選定してた. 一方, 交渉学等において, ≪賛成≫や≪反対≫を“(利害)関心”ではなく, “立場”と捉える考え方もある[10]が, 本研究では「真の関心」の分類軸として, ≪賛成≫や≪反対≫も含まれるものと考えている.

なお, 読み手による解釈のばらつきの傾向と類型化の検討, あるいは, それを軸とした自動分類を行う際には, 与えられた「疑問型表現文」一文に対して, 上記の八種類の“関心”のうち, いずれか 1 つを付与する必要がある. 以降, これらを“関心タグ”と呼ぶ.

表-1 疑問型表現文の分類軸

コード	内容	意　味
A	質問	疑問点やわからない点への問い.
B	疑い	うたがうこと. 怪しいと思うこと. ＜疑念, 不審を含む＞
C	確認	たしかめること. ＜同意を求めることを含む＞
D	要求	必要, また当然の権利として 強く求めること. ＜依頼, 提案を含む＞
E	不満	十分に満たされていないと思うこと. 満足しないこと.
F	懸念	気にかかって不安に思うこと. ＜不安, 心配を含む＞
G	賛成	人の意見や行動をよいと認めて, それに同意すること. ＜期待を含む＞
H	反対	ある意見などに対して逆らい, 同意しないこと. 否定的であること.
I	その他	A～Hまでの選択肢には当てはまらないもの.

4. 疑問型表現文における“関心”の解釈のばらつき

自由回答テキストを自動分類するにあたり, 意見の受け取り手である“読み手”の間においても, “関心”の解釈にどのような相違が生じうるのかを予め把握しておく必要がある. そこで, 本節では, 前節で定義した関心タグを被験者に付与してもらう実験を行い, 読み手の違いによって解釈がどのようにばらつくのか, その傾向を把握する.

(1) 関心タグの付与実験

まず, 前述の対象データ 1515 文からランダムに抽出した 215 文について, 関心タグの付与実験を行った(**表-2**). これを, タグ付与実験 I と呼ぶ

被験者には, 北西線の PI 手法を通じて市民から得られた疑問型表現文と, その前後の文を含めた回答全体を読むよう指示し, 読んだ文のうち, 疑問型表現文のみにタグを付与するように指示した. 回答は選択式とした. 選択肢(関心タグ候補)は, 言語論を参考にして設定した前述の八種類である. なお, これらに含まれないものは, “I. その他(自由回答)”と設定した.

(2) 実験結果の考察

タグ付与実験 I の結果について述べる.

- 対象とした 215 文に対し, 選択肢である 8 つのタグのうちいずれかが必ず付与され, 1 度も利用されなかったタグは存在しなかった.
- 被験者には, 最も当てはまっていると思われる関心タグを, 第一位から第三位まで割り当てるよう指示した. なお, 第二位と第三位の記入は任意とした. その結果, 被験者 10 名より, 全部で 194 パターンの回答結果が得られた. この 194 個のパターンは, ①一つの文章に対して A～I のタグをいずれか 1 つだけ（すなわち 1 位のタグだけ）を付与したもの, ②A～I からタグを 2 つ（すなわち 1 位と 2 位のタグを）付与したもの, ③A～I からタグを 3 つ（すなわち 1 位～3 位のタグを）付与したもの, によって構成されている. 詳細は庭田[11]を参照されたい.
- ここで, A～G のタグは高頻度（いずれも 80 件以上）で第一位に割り当てられていた. 一方, H に

表-2 タグ付与実験 I の概要

実験目的	①『疑問型表現文』内の“関心”の解釈の傾向 ②テキストの自動分類のための、『疑問型表現文』内の一般的な“関心タグ”選定
実験日時	2004年11月5日～11月16日
被験者	10名(東京工業大学交通・都市計画系研究室大学院生; 修士課程1年:8名,2年:2名)
実験内容	『疑問型表現文』を含んだ自由回答テキスト全体(回答者一人文の全意見)を読み、『疑問型表現文』のみに“関心タグ”を付与する。
回答方法	選択回答式(選択肢群から当てはまると思う順に1位～3位までの順位付け)

関しては，第一位への割り当て頻度は相対的に少ないものの（24 件），194 パターンある割り当て結果の中では，比較的，発生頻度が高い場所に位置している．

- 第一位－第二位の組み合わせで最も多かったものは A. 質問－F. 懸念（全体の 5%）であった．
- I. その他として，3 人の被験者により，215 文中の 7 文に対して，“いやみ”, “脅し”, “同意”, “理解不能”といった名称の 4 つの新しいタグが付与された．
- 解釈にばらつきが生じた文章には，“このような”, “その為に”, “そもそも”, “～でしょう．”, “～では．”, 等といた表現が含まれているという特徴が見られた．

以上で得られた知見は，次節で実施するタグ付与実験 II，及び，テキスト自動分類における，事前の基礎知見として活用する．

５．機械学習による自由回答テキストの自動分類

本節では，前節の実験を通じて解釈にばらつきが存在することが確認された北西線 PI 自由記述データを，人々の“関心”に基づいてどの程度の精度で自動分類できるかを検証するため，自由回答テキストデータに対して機械学習アプローチによる自動分類を行う．

(1) テキスト分類方法の概要

テキスト分類とは，予め設定された二つ以上のカテゴリに文書を分類するタスクのことを指す．一般に，テキスト分類では，文書を多次元のベクトルで表現する．例えば，“必要”, “道路”, “PI”, “行政”という 4 つの単語を“素性”（後述）として，その素性の出現をベクトルで表現すると，以下の 2 つの文書は**表-3** のように表記される．

文書 1：この道路は建設が必要である．

文書 2：現在行われている PI は必要だと思う．

ここで，各文書に対して，その文書が帰属するカテゴリのラベル y が与えられるとすると，テキスト分類のための概念装置（分類器）を割り当てる問題は，訓練データ $S=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\cdots,(\vec{x}_n,y_n)\}$ が与えら

表-3 テキスト文書のベクトル表現

	必要	道路	PI	行政
文書 1（$\overrightarrow{x_1}$）	1	1	0	0
文書 2（$\overrightarrow{x_2}$）	1	0	1	0

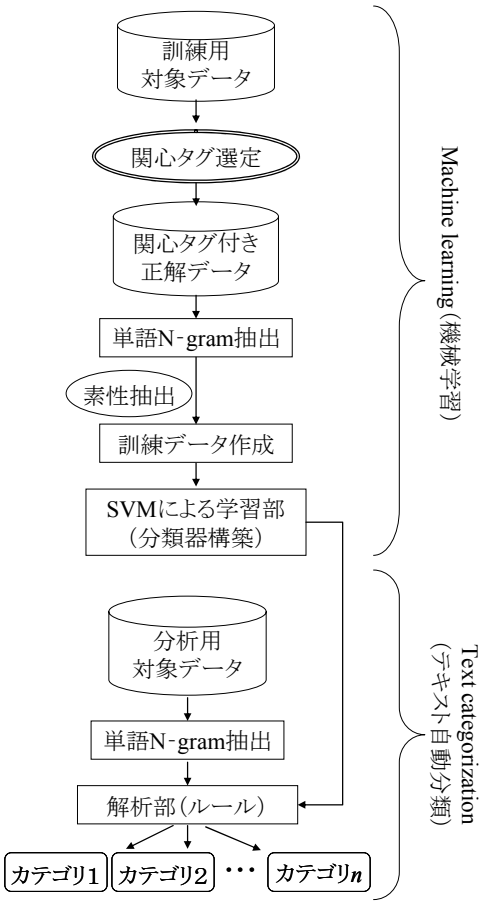

図-4 機械学習によるテキスト分類

れたときに，予測されたラベルが本当のラベルと異なる回数を最小化するような識別関数 $f(\vec{x})$ を求める問題と考えることができる．

本研究では，サポートベクトルマシン（SVM）[12]を用いて分類器を構築し，テキストの自動分類を行う．そのフローを**図-4** に示す．

(2) 関心タグ付き正解データの作成

前節のタグ付与実験 I をもとに，各自由回答テキストに対して，8 種類の関心タグのいずれかが付与された“関心タグ付き正解データ”を作成する．具体的には，タグ付与実験 I の回答において被験者の過半数が順位 1 位に回答した関心タグとそれが付与された文を関心タグ付き正解データとする．

(3) 素性抽出と訓練データの作成

次に，正解データから単語 N-gram による素性抽出を行う．単語 N-gram とは，一文を形態素（**表-4**）に分解後，N の数を変化させた任意の連続単語連鎖を使い，文章の特徴を取り出すことである（**表-5**）．言語学では，意味を担う最小の言語要素を“形態素”と呼ぶ．また，自然言語処理においては，1 つの文を，意

味を担う最小の言語要素＝形態素に分解する処理を“形態素解析”と呼ぶ．このとき，N = 1 は任意の形態素の集合を，N = 2 は，N = 1 に任意の形態素の 2 連鎖を追加した集合を，N = 3 は N = 2 に任意の形態素の 3 連鎖を追加した集合を…それぞれ指す．

形態素解析には茶筌 [13]というソフトを用いた．その際，北西線に特徴的な固有名詞 769 語（例．横浜環状北西線，PI，…）を予め辞書に登録した．

本研究では，N を 3 から 6 まで変化させた場合の影響について考察する．自由回答テキストの特性として，新聞記事とは異なり表現形式に個人差や表現のゆれなどが表れやすいため，それを文の特徴として抽出できるという利点から，N を変化させて比較検証を行う．

このような手順に沿って素性抽出を行い，タグ付与実験 I から得られた“関心タグ付き正解データ”を分類器の“訓練データ”に用いる．また，後述する実験 II から得られた“関心タグ付き正解データ”を“検証用データ”とする（**表-6**）．

表-4　形態素分解の例

（『神奈川新聞』を見たが，有識者委員会の傍聴はできるのか？）

表層語	基本形	読み	品詞-活用
『	『	『	記号-括弧開
神奈川新聞	神奈川新聞	カナガワシンブン	名詞-固有名詞-一般
』	』	』	記号-括弧閉
を	を	ヲ	助詞-格助詞-一般
見	見る	ミ	動詞-自立
た	た	タ	助動詞
が	が	ガ	助詞-接続助詞
、	、	、	記号-読点
有識者委員会	有識者委員会	ユウシキシャイインカイ	名詞-固有名詞-一般
の	の	ノ	助詞-連体化
傍聴	傍聴	ボウチョウ	名詞-サ変接続
は	は	ハ	助詞-係助詞
できる	できる	デキル	動詞-自立
の	の	ノ	名詞-非自立-一般
か	か	カ	助詞-副助詞／並立助詞／終助詞
？	？	？	記号-一般

表-5　N-Gram 分割の例

（“／”が区切りを表す）

N=2（bi-gram）； 『-神奈川新聞／神奈川新聞-』／』-を／を-見／見-た／た-が／が-，／，-有識者委員会 ／有識者委員会-の／の-傍聴／傍聴-は／は-できる／できる-の／の-か／か-?
N=3（tri-gram）； 『-神奈川新聞-』／神奈川新聞-』-を／』-を-見／を-見-た／見-た-が／た-が-，／が-，-有識者委員会／，-有識者委員会-の／有識者委員会-の-傍聴／の-傍聴-は／傍聴-は-できる／は-できる-の／できる-の-か／の-か-?
N=4； 『-神奈川新聞-』-を／神奈川新聞-』-を-見／』-を-見-た／を-見-た-が／見-た-が-，／た-が-，-有識者委員会／が-，-有識者委員会-の／，-有識者委員会-の-傍聴／有識者委員会-の-傍聴-は／の-傍聴-は-できる／傍聴-は-できる-の／は-できる-の-か／できる-の-か-?

(4) 機械学習

図-3 の機械学習部分では，予め答えが用意された訓練データを用いて，SVM による機械学習を行い，分類器を構築する．喩えて言えば，回帰分析において，未知パラメータである回帰係数を推定し，回帰モデルを同定する作業に相当する．

次に，この訓練データを用いて分類器の学習を行う．具体的には，入力された自動回答テキストに対して，各関心タグが分類先となる確率を機械学習させる（**図-5**）．ここで，本研究で分類器として採用した SVM は，訓練データを正例と負例に二値分離し，かつ正例と負例の間のマージン（＝訓練データから分離平面までの距離の最小値）が最大になるような超平面を求める学習器である．その過程は次のように定式化できる．

前田 [14] に倣い，二値分離のための線型 SVM について説明する．まず，識別関数 f を次のように定義する．

$$f(\vec{x}) \equiv \mathrm{sign}(g(\vec{x})) = \mathrm{sign}(\vec{w}\cdot\vec{x}+b) \qquad (1)$$

但し，$\vec{w}, b$ は未知パラメータ（ベクトル）である．ここでは，基準化のため次の仮定を置く．

$$\left|g(\vec{x})\right| = \left|\vec{w}\cdot\vec{x}+b\right| = 1 \qquad (2)$$

ここで，n 個の訓練用自由回答データ $\vec{x}_i (i = 1,\cdots,n)$

表-6　訓練データ・検証データの設定例

自由回答テキスト	関心タグ（y）	素性抽出（x）								
		なぜ	なぜこれ	・・・	農村地域	・・・	するのか	のか	か	か？
なぜこれまで つくらなかったのか？	A	1	1		0		0	1	1	1
なぜ農村地域に トンネルを計画するのか。	E	1	0		1		1	1	1	0

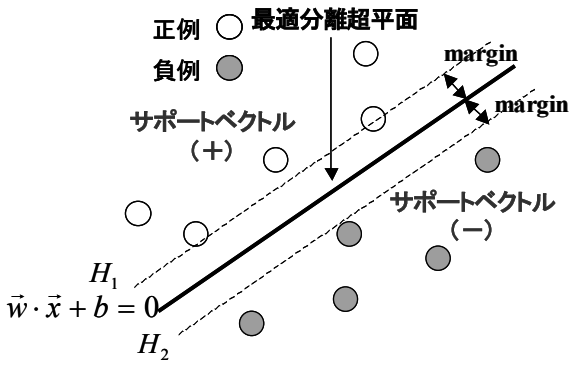

図-5　サポートベクトルマシンの概念図

が満たすべき条件として，次式を挙げる．

$$\forall i,\ g(\vec{x}_i)=\begin{cases}\geq 1 & \text{if 1番目のカテゴリに属する}\\ \leq -1 & \text{if 2番目のカテゴリに属する}\end{cases} \quad (3)$$

点 $\vec{x}_i$ から平面 $g(\vec{x})=0$ までの距離は $|g(\vec{x})|/\|\vec{w}\|$ で与えられることから，条件式(3) は，境界平面 $g(\vec{x})=0$ から距離 $1/\|\vec{w}\|$ までの範囲内，すなわち，二つの平面 H_1，H_2（$g(\vec{x})=\pm 1$）で挟まれる領域の間に学習パターン（訓練用自由回答データ）が存在しないことを意味する．ここで，$\vec{x}_i$ の属するカテゴリを変数 y_i で表し，次のように定義する（$\vec{x}_i$ の教師信号という）．

$$y_i=\begin{cases}+1 & \text{if 1番目のカテゴリに属する}\\ -1 & \text{if 2番目のカテゴリに属する}\end{cases} \quad (4)$$

すると，学習パターン $\vec{x}_i$ が満たすべき条件式(3)は，

$$\forall i,\ y_i\cdot(\vec{w}\cdot\vec{x}_i+b)-1\geq 0 \quad (5)$$

と書き直すことができる．そして，平面 H_1，H_2 の距離 $2/\|\vec{w}\|$ を最大にする f は，次の制約条件付き最大化最小化問題の解 $\vec{w}^*, b^*$ によって与えられる．

$$\text{Minimize } G(\vec{w})=\frac{1}{2}\|w\|^2$$
$$\text{Subject to } \forall i,\ y_i\cdot(\vec{w}\cdot\vec{x}+b)-1\geq 0$$

この制約条件付き最適化問題は，通常の Lagrange 未定乗数法を用いて解くことができる．テキスト分類のような非常に高次元の学習パターンで数クラスの分類を行う識別関数の最適化を行うような状況では，上述のマージン最大化基準が有効に働くことが分かっている [14]．

なお，これまで説明してきたように，SVM はそもそも 2 値分類器である．本研究では，8 つのカテゴリの分類を行うために，多値分類の代表的な方法である“one-against-all 法”（着目しているカテゴリとそれ以外のカテゴリに分類する 2 値分類器をカテゴリの個数だけ用意する方法）を適用する．

(5) 検証用データの作成

訓練データは，抽出された素性の正解データの出現を 2 値ベクトルで表したものである（**表-3**）．

また，自動分類結果の精度を検証するために，あらかじめ正解データを別途用意しておく必要がある．そこで，1515 文から前実験で用いた 215 文以外を除いた 1300 文から，ランダムに 288 文を抜粋し，同様のタグ付与実験を行った（タグ付与実験 II）．実験の概要は**表-7** の通りである．この実験では 182 文の正解データが得られており，訓練データの作成時と同様，被験者による分類結果とテキストの N-Gram 分割した結果を組み合わせて“検証用データ”を作成する．

(6) 適用

タグ付与実験 I で得られた学習データ 129 文を SVM の入力とし，任意の入力に対して各関心タグが分類先となる確率を学習させる．SVM の学習部については，ソフト Tiny-SVM[15] を用いた．

次に，学習が完了して構築された分類器を用いて，検証実験を行った．ここで，実験 II で得られた 182 文を検証用データとして用い，N-Gram 分割に関しては，素性を N=3，N=4，N=5，N=6 とした 4 つのパターンで行った．結果を**表-8** に示す．以下，分類器の精度に関して，“正解率”，“適合率”，“再現率”という 3 つの観点から考察する．

まず，正解率とは，実験結果の正解数を検証用正解データ数（=182）で除した値で与えられる．ここで正解数とは，N=3, 4, 5, 6 の各場合に対して，検証データで付与されたタグと分類器システムが出力したタグが一致したデータ数（＝**表-8** における各マトリクス (a)～(d) それぞれの対角要素の和）を指す．本研究では，最頻のカテゴリを選ぶ方法（最頻法）に基づき，関心タグのうち最も頻度の高い“A. 質問”の正解データ

表-7　タグ付与実験 II の概要

実験目的	テキストの自動分類における検証用正解データ作成のための『疑問型表現文』内の一般的な“関心タグ”選定
実験日時	2005年1月14日～1月19日
被験者	13名(東京工業大学交通・都市計画系研究室＋環境心理系研究室大学院生;修士課程1年:9名,2年:3名,博士課程3年:1名)
実験内容・回答方法	タグ付与実験 I と同様

表-8 SVM によるテキスト自動分類の結果

(a) N=3　　正解率= 59.3%

出力 / 正解	A	B	C	D	E	F	G	H	小計	再現率
A	49	0	0	1	1	0	0	0	51	96%
B	4	0	0	2	1	1	1	0	9	0%
C	1	0	0	0	0	1	4	0	6	0%
D	6	0	0	39	0	3	4	0	52	75%
E	10	0	0	3	4	1	1	0	19	21%
F	2	0	0	14	1	15	1	0	33	45%
G	0	0	0	4	0	2	1	0	7	14%
H	0	0	0	3	0	1	1	0	5	0%
小計	72	0	0	66	7	24	13	0	182	
適合率	68%	－	－	59%	57%	63%	8%	－		

(b) N=4　　正解率= 58.2%

出力 / 正解	A	B	C	D	E	F	G	H	小計	再現率
A	50	0	0	1	0	0	0	0	51	98%
B	4	0	0	2	2	1	0	0	9	0%
C	2	0	0	1	0	1	2	0	6	0%
D	7	0	0	40	0	3	2	0	52	77%
E	10	0	0	4	4	1	0	0	19	21%
F	2	0	0	17	1	12	1	0	33	36%
G	0	0	0	5	0	2	0	0	7	0%
H	0	0	0	3	0	1	1	0	5	0%
小計	75	0	0	73	7	21	6	0	182	
適合率	67%	－	－	55%	57%	57%	0%	－		

(c) N=5　　正解率= 58.2%

出力 / 正解	A	B	C	D	E	F	G	H	小計	再現率
A	50	0	0	1	0	0	0	0	51	98%
B	4	0	0	2	2	1	0	0	9	0%
C	2	0	0	1	0	0	3	0	6	0%
D	7	0	0	41	0	3	1	0	52	79%
E	10	0	0	4	4	1	0	0	19	21%
F	2	0	0	18	1	11	1	0	33	33%
G	0	0	0	6	0	1	0	0	7	0%
H	0	0	0	3	1	0	1	0	5	0%
小計	75	0	0	76	8	17	6	0	182	
適合率	67%	－	－	54%	50%	65%	0%	－		

(d) N=6　　正解率= 58.8%

出力 / 正解	A	B	C	D	E	F	G	H	小計	再現率
A	50	0	0	1	0	0	0	0	51	98%
B	4	0	0	2	2	1	0	0	9	0%
C	2	0	0	1	0	0	3	0	6	0%
D	7	0	0	42	0	3	0	0	52	81%
E	10	0	0	4	4	1	0	0	19	21%
F	2	0	0	18	1	11	1	0	33	33%
G	0	0	0	6	0	1	0	0	7	0%
H	0	0	0	3	1	0	1	0	5	0%
小計	75	0	0	77	8	17	5	0	182	
適合率	67%	－	－	55%	50%	65%	0%	－		

数（N=3 の場合は 49 文，N=4，5，6 の場合は各 50 文）をデータ総数 182 文で割った値，すなわち，全てに A というタグを付与した際の正解の割合をベースラインとみなす．その値は N=3 で 26.9%，N=4，5，6 で 27.5%であり，どのパターンにおいても，正解率がベースラインより高い結果となった．すなわち，最頻出語義を常に選択するベースラインモデルと比べて，適用した SVM に基づく機械学習分類器は，より高い精度を保持していることが明らかとなった．

次に，システムの正確性を判定する指標である適合率，並びに，網羅性の判定指標である再現率に基づいて検証を行う．まず，適合率は，タグ毎の正解数を分類器が出力した各タグに対する文の数で除した値によって定義され，**表-8** の各ケースにおいて，マトリクスの対角項の値を A～H それぞれの列の小計値で割った値として求められる．例えば，N=3 の場合におけるタグ A の適合率は 49/72≒68%となる．分類器の出力した数が 0 のタグに対しては，分母が 0 となるため適合率を定義することができない．一方，再現率は，タグ毎の正解数をタグ毎の学習正解データ数で除した値によって定義され，**表-8** の各ケースにおいて，マトリクスの対角項の値を A～H それぞれの行の小計値で割った値として求められる．例えば，N=3 の場合におけるタグ A の再現率は 49/51≒96%となる．

表-8 より，“B. 疑い”，“C. 確認”，“H. 反対”の各タグでは，N=3～6 のいずれの場合でも適合率が定義できない結果となっている（表中の“－”の記号）．この原因として，これらのタグの学習正解データ数が他のタグのデータ数よりも極端に少なく，分類器がそれらを正しく出力できなかったためと考えられる．一方，“D. 要求”タグでは，適合率がいずれの場合も 60%弱となっており，同時に，再現率が約 80%の高い値となっている．“F. 懸念”タグでも，適合率は平均して 60%を超えており，分類器の正確性の高さが確認されるが，再現率を見ると 33～45%となり網羅性はやや低い．“E. 不満”タグではそれが顕著で，適合率が平均 50%強であるのに対して再現率は平均で 20%強に留まっている．また“G. 賛成”タグでは，再現率が N=3～6 のいずれの場合でも 0%となっている．

以上より，当初仮定していた 8 つのタグには分類されず，疑問型表現文に含まれる市民の主要な“関心軸”は，“A. 質問”，“D. 要求”，“E. 不満”，“F. 懸念”，“H. 反対”の 5 つであることが明らかになった．

また，正解と出力結果とが異なるような状況（**表-8** の塗りつぶし部分）を確認すると，タグ付与実験において“B. 疑い”と判断されていた文が，“A. 質問”，“D. 要求”，“E. 不満”に誤判別される傾向があること，“C. 確認”と判断されていた文は“A. 質問”，“G. 賛成”に，また，“H. 反対”と判定されていた文は“D. 要求”に誤判別される傾向があることなども明らかになった．

(7) 誤判別要因に関する考察

分類結果の正誤例を**表-9** に示す．これより，

- “A. 質問”タグが付与される傾向があるのは，文末に“か？” という表現を持つ回答であること．
- “D. 要求”タグは，提案や賛成に近い “よい・のでは・よいのでは”といった表現を含む自由回答テキストに付与されやすいこと．
- “E. 不満”タグは，否定を表す“ない・ではないか”という文末表現を持つ文書に付与され易いこと．
- “F. 懸念”は，動詞の自立形である“なる・する”を含むテキスト，及び，否定形 “ないのでは”，あるいは，より否定を強くする“ますます”といった素性持つ文章に付与される傾向があること．

等が特徴として確認できる. このような誤判別の傾向に関する知見を蓄積して, 自動分類の精度向上を図る必要がある.

6．おわりに

本研究では，PI 活動を通じて得られた疑問型表現文における関心の所在を自動的に分類し, それらの素性，特に助詞や副詞，文末表現の構成に着目して，市民の関心軸の自動抽出を行った.

具体的には，SVM をベースとした自由回答テキストの自動分類の枠組みを構築し，様々な PI 手法を通じて集約されたテキストの「疑問型表現文」に特に着目して分析を行った．結果，回答者の “関心” の所在が複数個存在すること, すなわち, 質問・要求・不満・懸念・反対の 5 つの関心軸に自動的分類されることを明らかにした. また, 不満と解釈されたものは質問に, 同様に懸念と判断されたものは, 結果として要求に誤判別される傾向があるなど, 現状での自動分類の限界点も明らかにした. 分析の精度はまだ十分なものとは言い難く，本研究の成果を PI の現場において直接適用できる段階には至っていないが，今後，より大規模な訓練データ等を用いることによって, 分析の信頼性を向上させることができると期待している.

今後は, 精度向上のための訓練データの収集と適用や，前後の文章の構成との関連性の評価（文脈の影響検証）を行う必要がある. また, 今回は市民の “関心” といった観点からテキストの自動分類を行ったため, テキスト内容には言及しなかったが，実際には, “ルート”, “北西線”, “事業費”, “を - 通る” など，具体内容に関する素性によって自動分類される傾向が少なからずあることも分かっている. このような検討も行う必要がある.

謝辞

本研究の遂行に当たっては, 財団法人計量計画研究所大塚裕子氏, 東京工業大学精密工学研究所奥村学准

表-9　分類結果の正誤例

正解	正誤	出力	素性例	解析した自由回答テキスト原文
A	○	A	いつ,させる,させるつもり,つもりか,か？	構想段階はいつ頃までに完了させるつもりか？
A	○	A	どの,どのくらい,かかる,かかる,のか,か？	③計画段階が終わるまでにはどのくらいかかるのか？
C	×	A	いる,いるが,そうか,がそうか？,か？	川沿いと聞いているがそうか？
D	○	D	必要,では,ない,ので,でしょうか。	総合的な開発が必要ではないのでしょうか。
D	○	D	地下,が,よい,がよい,ので,よいのでは。	地下がよいのでは。
D	○	D	よいのでは,のでは,では,ではない,ではないか	丘は通してもよいのではないか。
F	×	D	のでは,のではないか,ではないか。	大気汚染の地域を拡張するだけになってしまうのではないのか。
G	×	D	のでは,ではないか,ない,のではないか。	一般道の交通事故も減るのではないか。
E	○	E	べき,べきだった,のでは,では,ない,のではないか,か。	横浜環状北線より先につくるべきだったのではないか。
E	○	E	ない,ないような,では,何のため,か。	トラックが通らないような料金設定では何のための高速道路か。
F	×	E	もし,として,して,してしまう,のではないか	もし北西線をつくったとして、普段から渋滞している生麦ジャンクション付近の渋滞を大きくしてしまうのではないか。
B	×	E	のでは,のではないか	おそらく市長は具体的なルートや住民（市民）の意見を十分に伝えられないままYesと言っているのではないか。
F	○	F	逆に,ますます,するの,するのでは,では？	逆に市が尾周辺がますます渋滞するのでは？
F	○	F	ない,ないような,,なるのでは,のでは,のではないか	利用もされないような道路になるのではないか。
F	○	F	ない,ないのでは,のでは	横並びでないと、たたき台案沿線以外の市民は関心を持たないのでは。
E	×	F	から,からだ,とは,られません,ませんか	渋滞がひどいのはこれまでの交通政策が間違っていたからだとは考えられませんか？
H	×	F	ない,ないので,ので,では,ないか！	必要ないのではないか！
H	×	G	ない,ないの,ので,ので,では,では？	いらないのでは？
F	×	G	ある,あるの,ので,では,ので,では？	トンネルー地下水の影響あるのでは？

教授，高村大也助教，乾孝司氏より，自然言語処理に関する様々な知識をご教授頂くと共に，分析上の留意点等，数多くのアドバイスを頂戴しました．この場をお借りして，感謝の意を表します．

参考文献

1) 高田伸二, 屋井鉄雄: アンケート自由記述による道路ニーズ・不満の把握手法の研究, 日本都市計画学会学術研究論文集, No. 35, pp. 571-576, 2000.
2) 針谷雅幸, 屋井鉄雄: 道路を対象としたアンケート自由記述の比較分析, 土木計画学研究・講演集, Vol.24, pp.525-528, 2001.
3) 矢嶋宏光: ワークショップ開催上の留意点, ワークショップ実例集－協働によるこれからの地域づくり, 全日本建設技術協会, 2006.
4) 坂野達郎, 永田典子: 発話プロトコルの分析による専門家と市民の総合計画に対する認識の差に関する研究－発話語の語彙的特長に着目して－, 計画行政, Vol.27, pp.43-51, 2004.
5) 内山将夫, 大塚裕子, 井佐原均: フェイスシートとの関係を利用した自由回答アンケートの分析, 信学技法, 2004.
6) 大塚裕子: 自由記述アンケート回答の意図抽出および自動分類に関する研究－要求意図を中心に－, 神戸大学大学院自然科学研究科博士論文, 2004.
7) 宮崎和人, 安達太郎, 野田春美, 高梨信乃: 新日本語文法選書 4－モダリティ, くろしお出版, 2002.
8) 仁田義雄, 益岡隆志: 日本語のモダリティ, くろしお出版, 1989.
9) 三省堂：大辞林 第二版.
10) Fisher, R. and Ury, W.: *Getting to Yes: Negotiating Agreement without Giving in*, Penguin Gooks, New York, 1983.
11) 庭田美穂: 自由回答の疑問型表現に着目した関心の抽出方法に関する研究, 東京工業大学大学院総合理工学研究科修士論文, 2005. (http://www.enveng.titech.ac.jp/yai/pdf/2004/niwata.pdf)
12) 金明哲, 村上征勝, 永田正明, 大津起夫, 山西健司: 言語と心理の統計, 岩波書店, 2003.
13) 形態素解析ソフト「茶筌」: http://chasen.aist-nara.ac.jp/
14) 前田英作: 痛快! サポートベクトルマシン－古くて新しいパターン認識手法－, 情報処理, Vol.42, pp.676-683, 2003.
15) 奈良先端科学技術大学：Tiny-SVM, http://chasen.org/~taku/software/TinySVM/

疑問型表現自由回答データを用いた社会資本整備に対する市民の関心の抽出方法に関する基礎的研究

福田大輔・庭田美穂・屋井鉄雄

本研究では，PI 活動を通じて得られた疑問型表現文における関心の所在を自動分類し，それらの素性，特に，助詞や副詞，文末表現の構成に着目して，市民の関心軸を抽出する方法についての検討を行った．具体的には，SVM をベースとした自由回答テキスト自動分類の枠組みを構築し，様々な PI 手法を通じて集約されたテキストの「疑問型表現文」に着目して分析を行った．分析の結果，回答者の“関心”の所在が複数個存在すること，すなわち，質問・要求・不満・懸念・反対の 5 つの関心軸に自動分類されることが明らかになった．

Fundamental Study of Abstracting Citizens' Interest toward Infrastructure Project by Using Interrogative Sentences of the Open-Ended Text Data

By Daisuke FUKUDA, Miho NIWATA and Tetsuo YAI

We conducted a fundamental study for automatically abstracting citizens' interest toward infrastructure project by using interrogative sentences of the open-ended text data from public involvement activities. We also focused on features of the text, particles, adverbs and expression of sentence end. A Support Vector Machine-based automatic abstracting system was developed and applied to interrogative sentences from citizens in various PI methodologies. It is found that the interest by citizens were automatically classified into the following four categories named “Asking”, “Claiming”, “Dissatisfaction” and “Objection”.
