自由回答の疑問型表現に着目した市民の関心の抽出方法に関する基礎的研究*

Abstracting Citizens' Interest toward Project from Open-Ended Text Focusing on Interrogative Sentence *

庭田美穂*・福田大輔**・屋井鉄雄***

By Miho NIWATA*, Daisuke FUKUDA** and Tetsuo YAI***

1．はじめに

事業の計画段階から住民の声を聴き，計画策定に反映させる PI の代表的な手法として，アンケート調査，FAX，電話，はがき等が挙げられる．特に近年，住民意見を直接表した自由回答方式が注目され，自由記述アンケートから道路に対するニーズや不満度を抽出する研究[1), 2)]等も行われている．PI では，どの手法を用いても，自由意見は最終的に「自由回答テキスト」となり，住民への公表や，計画プロセスにおける重要なデータとされる．しかし，それらの整理・分類法は未だ確立されていない．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，現在までに 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住民の意図や関心に注目した分類はなされていない．しかし，PI においては住民の関心の所在を明確にし，それを行政だけでなく住民全体が認識することが重要であり，そのためには，住民側から出される大量な自由意見を計画プロセスの中で埋没させないよう，各意見に秘められた人々の“関心”が何であるのかを効果的に抽出・分類する方法の確立が必要である．また，PI 手法を導入している行政側にとっても，膨大な意見を短時間で正しく理解し，整理した結果を的確に市民にフィードバックすることが求められている．

そこで，本研究では自然言語処理技術を援用し，「自由回答テキスト」に含まれる住民の関心を軸にしたテキストの自動分類を行う．特に，関心の相違が顕著である『疑問型表現文』に着目し，分類された結果から，『疑問型表現文』による意見表明における住民の関心の所在を明らかにする．さらに，意見を受取る側にとっての影響を知るために，読み手による関心の解釈の傾向を明らかにする．

*Keywords: 自由回答テキスト, PI, 自動テキスト分類

*非会員　株式会社インテリジェンス

**正会員　東京工業大学大学院理工学研究科土木工学専攻

（目黒区大岡山 2-12-1-M1-11　Tel: 03-5734-2577　Fax: 03-5734-3578）

***正会員　東京工業大学大学院総合理工学研究科人間環境システム専攻

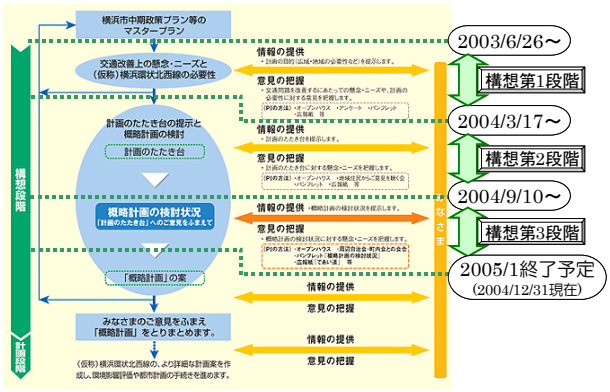

図-1　（仮称）横浜環状北西線－構想段階（HP より）

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

（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，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ており，具体的な手法として，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 段階へと進行している（2004 年 12 月 31 日現在）．

(2) 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（2003.6.26～2004.3.16）と②第 2 段階（2004.3.17～9.9）の自由回答テキストのうち，『疑問型表現文』（後述）を含む意見を対象データとする．それぞれの期間における疑問型表現文の数は，①が 934，②が 581 で，合計 1515 文となっている．

3．使用語句の定義

(1) 疑問型表現文の定義

本研究では，疑問型表現文を【言語論の“疑問のモ

ダリティ”＋その省略形, および疑問符】と定義する. 疑問のモダリティとは**図-2** に示すものであり,「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」, 文末に「～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, 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「何？」「いつ頃？」を含めている.

(2) 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[4]」と定義される. 本研究では, それら関心の所在を表す〈質問〉〈疑い〉〈提案〉〈不満〉などを“関心”と定義する. 読み手による解釈のばらつきの傾向と類型化, それを軸とした自動分類においては, 与えられた疑問型表現文一文に対し, “関心”を一つ付与する. よって以降, これらを“関心タグ”と呼ぶ.

4. 解釈のばらつきに関する検討

(1) 関心タグの付与実験

まず, 前述した対象データ 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（タグ付与実験 I）を行った（**表-1**）.

被験者には, 北西線の PI 手法によって個々の住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み, そのうちの疑問型表現文にのみタグを付与するように指示した. 選択回答式とし, 選択肢(関心タグ候補)は言語論を参考にして設定した, A. 質問, B. 疑い, C. 確認, D. 要求, E. 不満, F. 懸念, G. 賛成, H. 反対の 8 つ, これに含まれないものは I. その他(自由回答)とした.

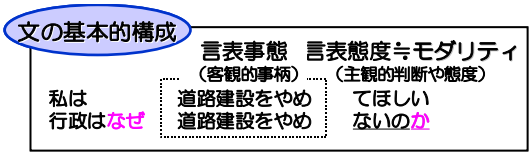

図-2　日本語文章の基本構成[3]

表-1　タグ付与実験 I の概要

実験目的	①解釈の傾向把握②関心タグの一般的付与
実験日時	2004年11月5日～11月16日
被験者	10名（東工大交通･都市計画系研究室学生）
回答方法	選択回答式（選択肢群からの1位～3位の順位付け）

(2) 考察

実験の結果, 選択肢の 8 つのタグがそれぞれに付与され, 1 度も利用されなかったタグは存在しなかった. また, I. その他として, 3 人の被験者により, 215 文中 7 文に“いやみ”“脅し”“同意”“理解不能”の 4 つの新しいタグが付与された. また, 全 194 パターンの回答が得られた. 解釈にばらつきがある文章には, “このような” “その為に”“そもそも”“～でしょう. ”“～では. ”という表現が含まれているという特徴があり, 読み手がこの計画がどのようなものか予め理解していなければ解釈できないこと, すなわち, 疑問型表現文だけでなく, 前後の文の捉え方も影響することが示唆された. 選択肢の組み合わせで最も多かったものは A. 質問－F. 懸念で, 全体の 5%であり, 解釈にも階層が存在することもわかった.

5. テキスト自動分類

(1) 方法論の概要

自由回答テキストの分類および関心の抽出の自動化を実現するため, 機械学習による自動分類を行う.

まず, 前節のタグ付与実験 I をもとに, 8 つの関心タグがそれぞれ付与された関心タグ付き正解データを作成する. ここから単語 N-gram 抽出を行い, 各タグに特徴的な表現を取り出して教師データとする.

次に, 訓練データを用いて分類器（サポートベクトルマシン：SVM）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学習させる（**図-3**）.

分類器として採用した SVM は, 訓練データを正例と負例に分離, かつ正例と負例の間の margin(=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(**図-4**).

まず, 識別関数 y_i を次のように定義する.

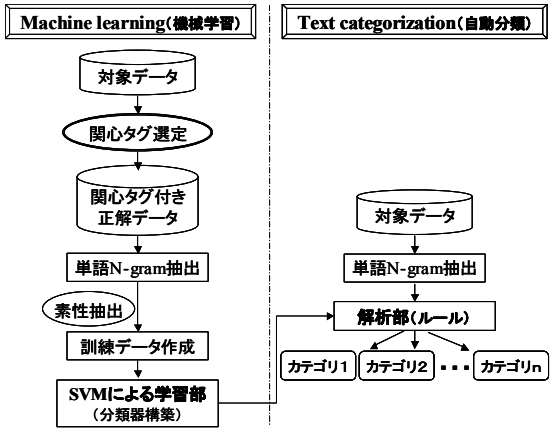

図-3　機械学習によるテキスト分類

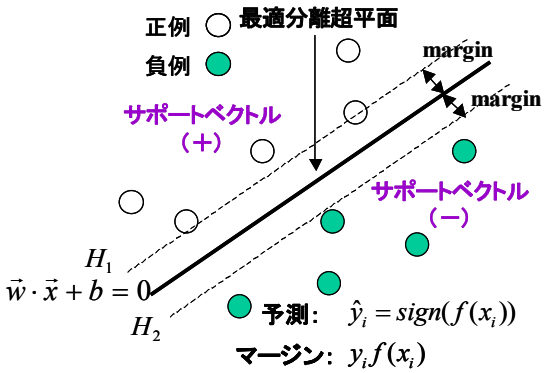

図-4　サポートベクトルマシンの概念図

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

ただし，基準化のために次の仮定をおく．

$$|\vec{w}\cdot\vec{x}_i + b| = 1 \tag{2}$$

ここで，n 個の学習パターン $\vec{x}_i (i=1,\cdots,n)$ について，点 $\vec{x}_i$ から平面 $f(\vec{x})=0$ までの距離は $|f(\vec{x})|/\|\vec{w}\|$ で与えられることより，境界平面 $f(\vec{x})=0$ から距離 $1/\|\vec{w}\|$ の範囲内，すなわち，二つの平面 H_1，H_2（$f(\vec{x})=\pm 1$）で挟まれる領域の間に学習パターン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 g_i で表し，次のように定義する（$\vec{x}_i$ の教師信号という）．

$$g_i = \begin{cases} +1 & \text{if カテゴリに属する} \\ -1 & \text{if カテゴリに属さない} \end{cases} \tag{3}$$

$\vec{g}_i = (g_1,\cdots g_n)$ とすると，学習パターン $\vec{x}_i$ が満たすべき条件は $g_i(\vec{w}\cdot\vec{x}_i+b)-1\geq 0$ と表される．すると，平面 H_1，H_2 の距離を最大にする y_i は，次の制約条件付き最適化問題の解 $\vec{w}, b$ によって与えられる．

$$\text{Minimize } G(\vec{w}) = \frac{1}{2}\|w\| \tag{4}$$
$$\text{Subject to } \quad g_i\cdot(\vec{w}\cdot\vec{x}+b)-1\geq 0$$

SVM はそもそも 2 値分類器であるが，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）を適用している．また，SVM の学習部については，ソフトウェア Tiny-SVM[5]を用いた．

(2) 関心タグ付き正解データ作成

タグ付与実験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．実験で用いた文数と正解データ数を PI の実施時期別にみると，①66/114，②63/111，合計 129/215（正解データ数/実験数）となる．

(3) 素性抽出

正解データから単語 N-gram による素性抽出を行った．単語 N-gram とは，一文を形態素ごとに分解後，N=1～6 まで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．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本抽出方法を用いた．形態素解析には茶筌 [6]というソフトを用いたが，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．この例として，“F. 懸念”のタグが付与された“**北西線ができることにより，インターチェンジ周辺が混むのでは？**”という文は N=6 抽出を行うと，“**北西線/北西線が/北西線ができる/北西線ができること/北西線ができることに/北西線ができることにより/･･･/混むのでは/混むのでは?/･･･/のでは?/で/では/では?/は/は?/?**”となる．ここで，N=6 までの変化をみたのは，疑問型表現文の一つの特徴である文末表現を考慮できると考えたためである．

(4) 訓練データ及び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-2**）．

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-3** 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，これを検証用データとする．

(5) 分析結果

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．分類実験は，実験

表-2　素性及び訓練データの例

	関心タグ	素性例									
		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
なぜ農専地域にトンネルを計画するのか。	E	1	0		1		1	1	1	0

表-3　タグ付与実験 II の概要

実験日時	2005年1月14日～1月19日
被験者	13名（東工大交通･都市計画，環境心理系研究室学生）
回答方法	選択回答式（選択肢群から1位～3位の順位付け）

II で得られた 182 文を検証用データとして行った．実験パターンは，素性を N=3，N=4，N=5，N=6 の 4 パターンで行った（**表-4**）．

正解率とは，実験結果の正解数(以下，正解数)/検証用正解データ数を表し，適合率(Precision)とは正解数/学習正解データ数，再現率(Recall)とは正解数/実験によりシステムが出力した各タグに対する文数を表す．ここで，最頻のカテゴリを選ぶ方法〈最頻法〉により，関心タグのうち最も頻度の高い“A. 質問”の正解データ数 41 文をデータ総数 129 文で割った値をベースラインとみなす．ベースラインは 0.395 であり，本実験のどのパターンでも正解率はそれ以上の値となり，ベースラインより高い精度となったことがわかる．それぞれの適合率と再現率をみると，B. 疑い C. 確認 H. 反対の各タグでは検証用データの適合率が 0 となっている．この原因として，学習正解データが他のタグのデータ数よりも極端に少なかったことが考えられる．D. 要求タグでは適合率が平均して 0.7，E. 不満では 0.6 を超えているが，同じく再現率をみると 0.6，0.3 とやや低い．E. 不満タグではそれが顕著で，適合率 0.6 に対し再現率 0.2 弱である．また，G. 賛成タグでは実験結果の正解率が 0 であることがわかる．

表-4　SVM による自動分類の結果

	N=3		N=4		N=5		N=6	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
A.質問	0.729	0.843	0.712	0.824	0.721	0.863	0.717	0.843
	(43/59)	(43/51)	(42/59)	(42/51)	(44/61)	(44/51)	(43/60)	(43/51)
B.疑い	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0/0)	(0/1)	(0/2)	(0/3)	(0/4)	(0/5)	(0/6)	(0/7)
C.確認	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0/0)	(0/6)	(0/0)	(0/6)	(0/0)	(0/6)	(0/0)	(0/6)
D.要求	0.689	0.596	0.705	0.596	0.698	0.577	0.714	0.577
	(31/45)	(31/52)	(31/44)	(31/52)	(30/43)	(30/52)	(30/42)	(30/52)
E.不満	0.600	0.158	0.600	0.158	0.600	0.158	0.600	0.158
	(3/5)	(3/19)	(3/5)	(3/19)	(3/5)	(3/19)	(3/5)	(3/19)
F.懸念	0.632	0.364	0.647	0.333	0.714	0.303	0.667	0.242
	(12/19)	(12/33)	(11/17)	(11/33)	(10/14)	(10/33)	(8/12)	(8/33)
G.賛成	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0/3)	(0/7)	(0/1)	(0/7)	(0/0)	(0/7)	(0/0)	(0/7)
H.反対	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0/0)	(0/5)	(0/0)	(0/5)	(0/0)	(0/5)	(0/0)	(0/5)
正解率	0.593		0.582		0.582		0.588	
	(108/182)		(106/182)		(106/182)		(107/182)	

表-5　分類結果の正誤例

正解	正誤	出力	素性例	解析データ
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,のでは,のでは？	いらないのでは？
F	×	G	ある,あるの,のでは,のでは？	トンネルー地下水の影響あるのでは？

⑹ 考察

分類結果の正誤例を**表-5** に示す．これより，A.質問タグが付与される傾向として，“か？”D. 要求タグでは提案や賛成に近い“よい・のでは・よいのでは”，E.不満タグでは否定を表す“ない・ではないか”，F.懸念では動詞の自立形である“なる・する”と否定形“ないのでは”，より否定を強くする“ますます”といった素性が特徴的である．また，質問では疑問詞が素性として多い．どのタグでも共通の“～では”については，前後の文章構成や連鎖になっていない他の助詞や副詞との組み合わせの頻度傾向から，関心抽出をより深めることが出来ると考える．

６．おわりに

本研究では，機械学習による自動分類手法を用いて，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，文末表現の構成により，関心を抽出した．今後は入力データ数による精度向上，前後文との関係等の課題が残されている．

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:道路を対象としたアンケート自由記述の比較分析，東京工業大学土木工学科卒業論文, 2001.
3) 仁田義雄，益岡隆志: 日本語のモダリティ，くろしお出版，1989.
4) 三省堂：大辞林 第二版.
5) 奈良先端科学技術大学：Tiny-SVM, http://chasen.org/~taku/software/TinySVM/
6) 形態素解析ソフト「茶筅」：http://chasen.aist-nara.ac.jp/