対話型意見収集システムの評価方法の検討

大塚裕子[1]　乾孝司[2]　鈴木泰山[3]　伊藤裕美[1]　丸元聡子[1]　奥村学[2]
[1]計量計画研究所(IBS)　[2]東京工業大学　[3]株式会社ピコラボ

1. 背景と目的

本研究で開発している対話型意見収集システム（IOCS: Interactive Opinion Collection System）は基本的に、ユーザーからの入力情報に対して、1）その理由を尋ねる、2）その詳細を尋ねる、という二つの行為を繰り返し行っている。この繰り返しを行う過程で、ユーザーの初期入力の意見の理由や根拠、あるいは意見の背後にある関心事や懸念を掘り下げることを目的としている（丸元他 2008, 大塚他 2007）。

IOCS は、現在、市民参画型公共事業（パブリック・インボルブメント：ＰＩ）における支援ツールとして開発されているため、扱っている知識については領域依存性も高い。しかし、そのような領域依存性に関わるシステムの目的や利用する知識をメタ情報とみなした上で、システムの評価項目について検討し、その知見を蓄積していくことは、領域依存性の壁を乗り越えることにつながる。したがって、言語処理分野として共有するに値する知見であると考えている。

対話型のシステムについては、近年、CGM（consumer generated media）への関心により、自然言語処理の分野でもインタビューエージェント（鳥澤 2007）、能動的質問生成（伊藤・荒木 2007）による知識獲得の研究が注目されている。いずれも、目的を知識獲得として浅い知識により質問を生成するアプローチをとっている。このアプローチは、彼らの研究が領域非依存性を目指すのに対し、本研究は領域依存であることの違いはあっても、対話型の情報収集システムとして、何が評価されるべきかという課題は共通していると考える。

本稿ではわれわれが実装および実験した IOCS によって収集できた意見の分析をもとに、システムの目的を踏まえた上で、何を評価指標とすべきかという問題に関する情報を共有することを目的に、評価に関する課題について述べる。

2. 意見収集の方法と課題

ここでは、まず、なぜ対話型の意見収集システムが必要なのかという前提について検討しよう。

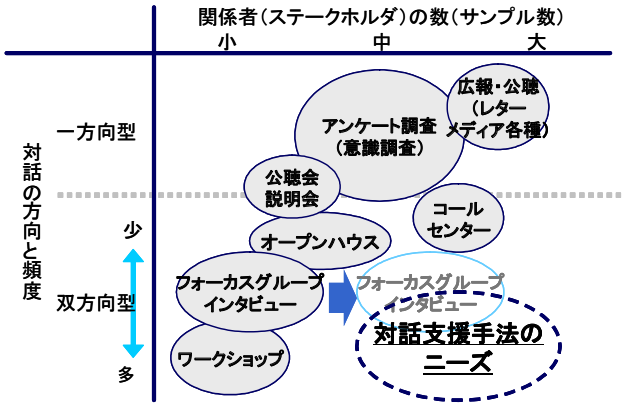

図1　一方向型および双方向型の情報伝達・収集方法の整理と課題

計算機による知識や情報の収集だけでなく、一般に知識や情報を収集する従来の方法について概観する。図１に示すように、情報収集のプロセスにおける対話（やりとり）の頻度を縦軸に、プロセスに関与する人々を横軸に設定し、各情報収集あるいは調査手法を配置すると、ワークショップやフォーカスグループインタビューなどの直接対話手法は左下に、アンケート調査などの一方向方の情報収集は右上に位置づけられる。フォーカスグループインタビューは、対話参加者をサンプリングやリクルーティングによって選ぶという点で、図１中の薄い矢印で示すように潜在的には大きな母集団の意見を得ていることになるが、サンプリングの前提として「情報収集すべき対象者には、どのような関係者が、どのように分布しているのか」などが明らかになっている必要がある。

多数多様な意見に対処するだけであればアンケート回答のテキスト分類（山本 2006）でもよい。しかし、アンケートは前述のとおり対話形式を成しているが、この相互作用が一回限りであるために対話である効果、すなわち「相手の反応に応じて行動（発話）する」という動的な性質と、その繰り返しの効果である「反応が不明であれば何度か確認する」という、人が本来行なっている行為の利点が反映されない。

これらの問題点から、図１に示すように、情報の伝達や収集などコミュニケーション（広い意味での対話）の支援に求められるのは、複数回のやりとりがある双方向型の対

話であり、かつ、関与する関係者の数が多いという点線で囲まれた要件を有する対話手法であることが明らかになってきている。IOCS はこのような問題意識のもとに開発実装された。

3. 評価項目の検討

本稿では、IOCS の評価を、１）対話型意見収集システムであることの評価と、２）市民参画型公共事業プロセス支援ツールとしての評価の二つに分けて検討する。１）については、一般的な言語処理技術による知識獲得、情報収集の対話型システムに共通する側面であると考える。２）については、「領域依存性」として説明される側面である。なお、これらの検討はすべて（丸元他 2008）で述べた、ある自治体でのシステム試行実験の際に得られた 36 人分の対話ログ 214 ターンの対話データを対象としている。

3.1 対話型意見収集システムであることの評価項目

（鳥澤 2007）にも述べられているように、対話エージェントの評価は、何を評価すべきかが確立されていないため非常に難しい。ここでは、対話のプロセスとコンテンツとを区別し、評価の検討を行う。

対話のプロセス（評価の観点：対話の流れが自然か）

対話のコンテンツ（評価の観点：得たい情報が得られているか）

また、プロセスとコンテンツの区別は、「対話システム自体の評価」と「対話システムの使用によって得られる効果の評価」の区別にも相当する。

A）ユーザーにとって対話が自然であるか、妥当であるか（システム自体の評価）

対話型のシステムにおいて、対話が自然か、妥当かということは、システム自体の評価に関わる評価項目である。この評価項目は、さらに、ア）質問は自然であり妥当であるかという談話分析的観点と、イ）ユーザーは対話を続けようとするかというユーザーの行為分析的観点に分けることができる。

ア）質問は自然であり妥当であるか

この観点の評価について、（鳥澤 2007）では、自動生成した質問文の文脈依存性を４段階の基準項目に分け、被験者にそれぞれの項目をラベリングさせることで評価を行っている。また、（伊藤・荒木 2007）は、生成した質問のうち、適切な質問文でないものをシステムのユーザーに削除してもらうという質問淘汰の仕組みをシステムに作ることによって、この問題に対処している。

本研究では、質問生成は予め用意された質問テーブルの中から、ユーザーの入力文のパタンに合わせて出力される。しかし、（丸元他 2008）でも示したように、質問設計自体はシンプルに、質問文にあわせた作りこみは行っていないため、表現や内容としても妥当でない問い返し表現が選択される場合がある。この評価について、次のようなテストを行った。

研究開発に携わっていないある一人の被験者（30 代女性、大学院博士後期課程在籍中、社会人経験有）によって、対話ログ 214 ターンすべてを対象に、「ユーザー入力$_n$－システム出力$_n$－ユーザー入力$_{n+1}$」という発話の三つ組みに対して、システム出力$_n$である問い返しが、ａ）問い返しの表現として適切か、ｂ）問い返しの内容として適切かという観点で三段階評価（1 よい、2 どちらともいえない、3 わるい）を行った。この際、システムの問い返しの前後のユーザー入力$_n$およびユーザー入力$_{n+1}$が変化したこと、あるいは変化しなかったことに着目し、システム出力である問い返しの妥当性を評価する。このテストで意図しているのは、対話型システムにおけるミクロな隣接ペアの妥当性、すなわち質疑応答の妥当性の評価である。この結果、表１のような結果が得られた。

表1 問い返しの表現と内容の適切さ判断

	1:よい	2:どちらともいえない	3:わるい
a) 問い返し表現	115	22	68
b) 問い返し内容	97	41	67

問い返しの表現・内容ともに「よい」とされたのは下線部の問い返しである。（Ｕ：ユーザー，Ｓ：システム）

#1 Ｕ:関係機関との協議を行う必要がある。
Ｓ:確かに．．．とくに、どんな点から必要と感じたか、〇〇さん、詳しく聞かせてください
Ｕ:費用負担

ユーザーの発話を受けて、意見の根拠や新たな入力を促す問い返しが表現、内容ともに「よい」と評価されている。一方、問い返しの表現・内容ともに「わるい」とされたのは下記の例である。

#2 Ｕ:評価項目が判りません。
Ｓ:一つずつ確認させてください。お聞きしたのは、おおまかに滑走路増設案に関するご意見ということでいいですか？
Ｕ:違います。
#2 Ｕ:県経済の発展ということです。
Ｓ:なるほど．．．長期需要についてのご意見ですね。〇〇さんは、長期的な需要が高まらないことを懸念されていると受け取りました。〇〇さん、長期的な需要が高まらない場合、どのようなことが気になるか、具体的に教えてください。

U:そのようなことは言っていません。

いずれも、直前のユーザーの発言を受けていないことが見て取れる。この評価の要因は、そのまま、(丸元他 2008)で指摘する質問設計の問題点に結びついている。

#2,3 の例は、質問設計上の典型的な問題事例であるが、これにより、対話を途中終了したと思われるケースも少なくなかった。これに関連して、次にユーザーが対話を続けるか否かという観点での評価を示す。

イ) ユーザーは対話を続けようとするか

IOCS は、ユーザーの希望で途中終了を許容する設計となっている。表2に示す2種類の対話途中終了には終了時における対話ログの利用許諾に対する判断の明示/非明示の区別がある。

表2 対話完了状況ごとのやりとりパタンの違い

	対話完了	対話途中終了	対話途中終了[*1]
アクセス数(人数)	13	8	15
ターン数[2]	104	57	53
ターン数平均	8	7.125	3.533
理由を尋ねる問い返しの数	21	6	5
詳細を尋ねる問い返しの数	29	26	17
確認a[3]への肯定	―	―	―
確認 a への否定	0	5	0
一つずつ確認	23	18	12
確認b[4]への肯定	20	10	9
確認 b への否定	3	5	3
促し	7	4	5
問2	11	3	4
お礼	13	―	―

表2に示すように、36 人のアクセスのうち、対話を完了した人は全体の3割という結果になった。対話完了者は、確認ステップ(確認 b)で肯定をすることが多い。これは、システムに対して「ユーザーの意見を理解した/受け止めた」と考えたことの現れとみなすことができる。

システムの途中終了者には、表2のように特徴的な傾向が見られた。対話完了者に比べて、1回のやりとりあたりのターン数が少ない、システムからの確認に「いいえ/違います」といった否定を表明する割合が高い、システムからの確認ステップ(確認 b)に応えず対話を辞めてしまう等である。確認への否定については、システムの理解度やユーザーのシステム使用満足度を測る上で重要な指標と考えられる。

B) 得たい情報が得られているか(システムの効果の評価)

(伊藤・荒木 2007)は、知識の獲得率を下記(1)のように定義し、能動的質問生成による web からの知識獲得率をシステムユーザーごとに評価している。

$$\text{知識獲得率}(\%) = \frac{\text{知識獲得数}}{\text{システム質問数}} \times 100 \qquad (1)$$

本研究の対話型意見収集システムでの、「得たい情報」とは、市民参画型の公共事業計画において重視される市民の「計画に対する関心や懸念(インタレスト)」であるため(丸元他 2008)、インタレストの獲得率は重要である。本研究のシステムのパフォーマンスを伊藤らの式にあてはめてインタレスト獲得率を算出すると、59.02%という獲得率が出る。[5]

$$\text{ｲﾝﾀﾚｽﾄ獲得率}(\%) = \frac{\text{ｲﾝﾀﾚｽﾄ獲得数}}{\text{システム質問数}} \times 100 \qquad (2)$$

しかし、この値は本研究のシステム評価にはあまり意味がない。なぜなら、本研究では(2)式が示す「一つの質問に一つの情報(ｲﾝﾀﾚｽﾄ)が得られた場合に 100%のパフォーマンスとしてカウントする」という指向性がないからである。市民参画における対話型意見収集システムでは「繰り返し問い返すことによって、インタレストを得ること」を目的としている。したがって、システムからの問い返し数が多くても、一つでもやりとりの中でインタレストを得ることができれば、それは 100%の成功であると考えるのである。この観点からのｲﾝﾀﾚｽﾄ獲得率を定義した式が(3)である。

$$\text{ｲﾝﾀﾚｽﾄ獲得率}(\%) = \frac{\text{ｲﾝﾀﾚｽﾄが得られたユーザー数}}{\text{ユーザー総数}} \times 100 \qquad (3)$$

また、インタレストの獲得において、もうひとつ重要なことは、問い返し(2回目以降の問い)によって得られるインタレストをシステムの効果の評価として重視するということである。なぜなら、最初の問いによって得られたインタレストは自由回答型アンケートのように、一方向型の情報収集システムでも獲得できる可能性があるからで

[1] 対話ログの利用許諾に対して判断不明なユーザーの回答

[2] 1 ターンは、ユーザー入力$_n$−システム出力$_n$

[3] 確認aとは、詳細を尋ねる問い返しの一部である「なるほど…××についてのご意見ですね。○○さんは××を懸念されていると受け取りました。」という確認

[4] 確認bとは、インタレストを推定した後の「一つずつ確認させてください」という確認

[5] システム質問数にカウントしたのは、理由を尋ねる問い返し+詳細を尋ねる問い返し+問いかけ2

ある。この観点から、インタレストの獲得率を示したのが図２である。

適合性の側面からは、問い返しによって得られるインタレストの獲得率が重要であるが、これはあくまでシステムによって推定されたインタレストである。再現性の側面からは、そのインタレストの確認時にユーザーが肯定した割合が確定されたインタレストの獲得率であるとみなすことができ、また、評価の項目としてももっとも重要であるといえる。今回の実験結果からは、72 件の推定インタレストのうち、確認ｂへの否定数を引いた 42 件が確定インタレストとなった。したがって、確定インタレストの獲得率は 58.3%である。

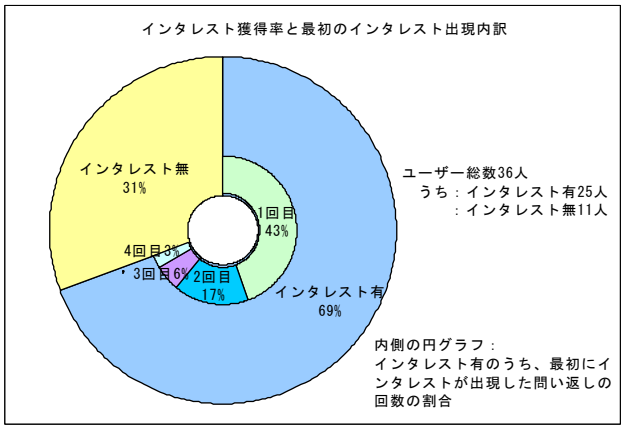

図2　インタレストの獲得率と出現内訳

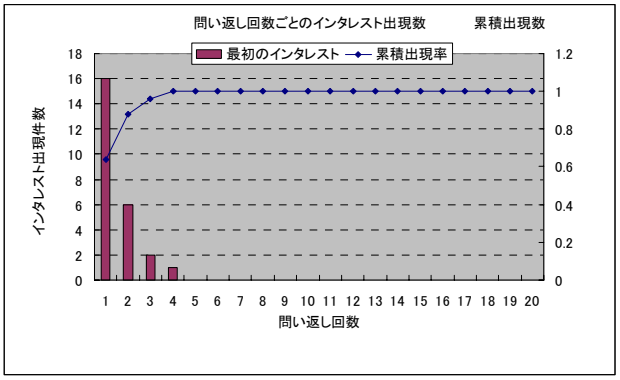

図3　問い返し回数ごとのインタレスト出現数

3.2 市民参画型公共事業プロセス支援ツールとしての評価

市民参画プロセスの支援ツールとしては、どのようなインタレストが取り出せたかということが評価指標となる。この評価項目は、領域依存性が高いため、今回はインタレストの種類が概ね把握できるよう図４を示すに留める。領域依存の部分については、（国交省 2001）で公開され、制度化されている公共事業評価の評価指標との比較も重要である。本研究の試行実験で対象としている計画中の指標と、（国交省 2001）が対象としている事後評価の指標では必ずしも一致しないと思われるが、比較によって計画体系における検討項目の整理に結びつくと考える。

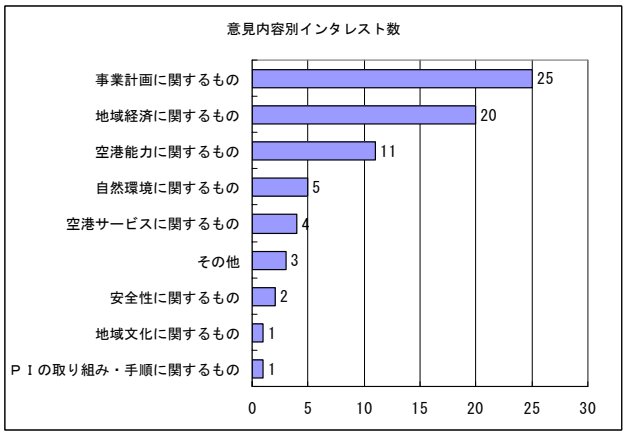

図4　獲得された内容別インタレストの数

4. まとめと課題

本稿では、対話型意見収集システムの実際のＰＩの現場における試行実験をもとに得られた意見の分析から、対話型システムの評価方法について検討した。現場ニーズに応えるシステムは、その評価についてもシステム実装の目的などに大きく依存する面がある。「得たい情報を獲得できたか」という評価指標は、その一つであり、本稿では「インタレストの獲得率」として議論した。

一方、対話型システムの評価として、個々の目的に依存しない「対話の自然さ」など共通に検討すべき項目もある。「自然さ」を評価する方法や評価項目には、まだ検討と議論の余地がある。

また、さらに重要なのは、「得たい情報を獲得」するための手段、道具としての「対話の自然さ」という、両指標の重みバランスをどう考えるかということである。これについては、今後の課題としていきたい。

参考文献:

伊藤慎吾・荒木健治，能動的質問生成を用いた対話メディアによる知識の獲得および提供，情処研報，Vol.2007,No.94, pp.121-126，2007.

国土交通省公共事業評価システム研究会，公共事業評価の基本的考え方，
http://www.mlit.go.jp/kisha/kisha02/13/130830_.html,2001.

丸元・鈴木・大塚・乾・奥村，空港計画における対話型意見収集システムの実装と課題，言語処理学会年第 14 回年次大会発表論文集，2008.

大塚・丸元・岩佐・鈴木・矢嶋・奥村・屋井，市民参画型道路計画における対話支援－対話型アンケートシステムの実装に向けて－，交通工学，Vol.42，No.2，pp.47-57，2007.

鳥澤健太郎，一般ユーザーにインタビューする対話エージェント，情処研報，Vol.2007,No.76，pp.25-30，2007.