自然言語処理　138－1
（2000．7．18）

コメント生成のための情報統合法の提案

佐藤 浩史　　松下 光範　　藤本 和則　　加藤 恒昭
NTT コミュニケーション科学基礎研究所

619-0237　京都府相楽郡精華町光台 2-4
{hiroshi, mat, fujimoto, kato}@cslab.kecl.ntt.co.jp

あらまし　ユーザの直観的な理解を支援する機能を持つ情報提供システムを実現するために，提示情報にコメントを付与する技術の研究を行っている．インターネットを情報源とした場合，その不完全性から情報を統合し信頼性を高める必要があり，また，その情報の信頼性をどう判断し扱うかが重要となる．システムは，信頼性が高くかつコメントの対象として適切な情報を選択し，さらにその信頼性をコメントに反映させなければならない．そこで，信頼性を考慮した情報統合法およびコメント対象の選択法を提案する．これは，情報の信頼性を可能性と確信度の2値で表現することを特徴としている．

キーワード　情報提供システム，コメント生成，情報統合，信頼性，概念階層

An Information Integration Method for Generating Adequate Comments

Hiroshi SATO, Mitsunori MATSUSHITA, Kazunori FUJIMOTO, Tsuneaki KATO
NTT Communication Science Laboratories

2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
{hiroshi, mat, fujimoto, kato}@cslab.kecl.ntt.co.jp

Abstract

This paper proposes an information integration method for information providing systems. This integration method handles the reliability of information, that is essential to generate *comments* from information on the Internet to help users intuitively understand the information and have favorable impressions of the system. The reliability in this method is represented by two factors: the possibility and certainty. The system uses the reliability to choose information that is reliable and proper for commenting purposes, and generates comments while reflecting the reliability of the information.

key words　Information Providing System, Comment Generation, Information Integration, Reliability, Concept Hierarchy

1 はじめに

情報化社会と言われるようになって久しい．特にインターネットは，その利用者数が著しく増加し，一般の人々にとっての生活情報源という地位を確固たるものにしつつある．しかしその情報は膨大であり，誰でもが欲しい情報を的確に獲得できるわけではない．そのため，この問題を解決する情報提供システムが求められており，その実現に必要な技術の研究が盛んに行われている [1, 2]．主な技術を挙げると次の2つとなる．

- 情報の検索 (Information Retrieval: IR)
- 情報の抽出 (Information Extraction: IE)

これらの技術は重要であり，研究がさらに進んでいけば，ユーザは必要な情報をかなりの程度入手することが出来るようになるかもしれない．しかし，システムがユーザに対し情報の所在（例えば URL）のみを教える，または抜き出した情報を羅列して見せるだけでは，ユーザが情報をすぐ理解できる保証はない．そこで我々は，IR, IE の技術に加え，

- 情報の直観的理解の支援

として，情報に対して簡単な供述－コメント－をシステムが独自に生成し，付与する技術の研究を行っている．これにより，インターネットを情報源とした直観的かつフレンドリな対話型情報提供が可能となる．

情報に付与するコメントは，情報の内容を直観的に示すだけでなくその質も反映させなければならない．そのためには，コメントの対象となる情報の信頼性を適切に扱うことが必要となる．本稿ではこれらを背景として，コメント生成のための情報の統合法および補完法について，特に情報の信頼性の扱いを中心に論じる．また，問題の理解を容易にするため，ドメインの例として天気予報案内を採り上げ，議論を進める．

コメント生成を含む天気予報案内のモデルは以下のようなものである．まず，要求された地域の気象情報をインターネット上の天気予報サイトから獲得する．この時，情報の信頼性を高めるために複数サイトから同地域の情報を獲得し統合を行うとともに，その統合情報の信頼性を算出する．その後に，独自の知識ベースを用いて統合情報と日常的な事象を結びつけ，コメントとして気象情報とともに提供する．天気予報案内におけるコメントとは，お天気キャスターが予報の最後に加えるひと言に相当する．例えば，「夕方から雨」という予報に対して加える「折り畳み傘が必要ですね」といったひと言である．

なお，天気予報サイトの情報は，予報の的中率はさておき，組織的に運営されているものが多く一般に信頼性が高いと考えられるが，サイトによって更新頻度に差があるため予報値が異なる場合がある．また気象業務法の改正により，以前は気象庁のみが行なっていた気象予報を，気象予報士の有資格者であれば各市町村の 24 時間以内の気象予報を行なえるようになったため，今後は情報の多様性が増し，今回提案するような信頼性を適切に判断する技術が必要になってくると思われる．

2 コメント生成

2.1 直観的な情報提供

一般の情報提供システムでは，獲得した情報の体裁を整えそれをユーザに表示するといった形態が多い．例えば，天気予報ならば「晴れ．最高気温 25 度」のような各種の気象情報を表示するし，外国為替情報ならば為替レートのチャートを表示する．これらはデータのみを求めているユーザにとっては必要十分な情報かもしれないが，一般のユーザにとっては情報を理解するのが困難な場合が多々あり，また，不親切な印象を受けることも否定できない．では，人間対人間での情報のやりとりはどうだろう．人が何か情報を得ようとして，誰かその情報について詳しい人に尋ねる場面を考えてみる．質問者がまず期待しているのはその情報に関する要約であり，その後に詳細を知るための説明を聞く，または詳しい資料を手にしようとするはずである．情報または資料だけを渡されるだけでは質問の回答としては不十分であるし，不親切だとさえ感じるだろう．そこで我々は，直観的かつフレンドリな情報提供のために，情報に対してシステムが独自に要約に類する付加的な情報を加えることを提案する．

付加情報の一つとして我々が考えているものは，提供する情報に対する簡単な供述，いわゆる「コメント」である．例えば，数値情報を直に羅列されてもその意味するところは捉えにくい．しかし，その数値の評価をコメントとしてひと言加えられれば，ユーザはより直観的に情報が理解できることになり，また同時にフレンドリな印象を持つだろう．コメントの形式としては，自然言語文を考えている．従って，ユーザとのインタフェイスは対話型となる．例えば，ユーザが京都府の気象情報を要求し，「晴れ．最高気温 20 度」を提供したとする．そこにコメントとして，「京都府は爽やかな天気ですね」を加える．これは天気をひと言で表しており，どのような天気なのかを直観的に理解しやすくなる．さらに，「外に出かけましょう」などと日常生活を絡めたコメントを加えれば，人間らしい親しみを感じさせることが期待できる．これが，我々の目指すシステムである．

2.2 必要な技術

このコメント生成を実現するためには，まず情報とコメントを結びつける知識ベースが必要なのは当然であるが，コメントの元となる情報がインターネット上のサイトから獲得したものであることを考えると，その情報の信頼性を考慮する必要がある．一般にインターネット上の情報は不完全なものであり，そこには間違いや矛盾，欠落が存在するからである．

コメントを自然言語文で生成するならば，信頼できる情報に対するコメントとそうでない情報に対するコメントでは，おのずとその表現に違いが出るはずである．つまり，「最高気温 30 度」という情報に対するコメントの内容として「真夏日になる」を生成した場合，元となる情報の信頼性に応じて，「真夏日になりそうです」や「真夏日になるかもしれません」などの表現を使い分けなくてはならない．この使い分けにより自然な発話になると同時に，信頼性の情報も直観的にユーザに提供できることとなる．この技術に関して，3章で述べる．

また，情報に欠落がある場合はその情報を推定し補完することを考えるが，一般に補完情報の信頼性は低く，その情報が要求されていたとしても，補完情報を提供することが必ずしも最良とは言えない．例えば天気予報の対象となる地域のように，情報対象が階層構造を持っていれば，上位（京都府）の情報が無くても下位（京都府北部／南部）の情報から推定できるが，推定した情報の信頼性が低いときはむしろ下位地域の情報をそれぞれ提供するほうが無難である．しかしコメントに関しては，ユーザが京都府の天気を求めている以上，可能ならば京都府に対して加えるほうが望ましい．つまり，「京都府の天気は？」という質問に対し，「京都府北部は雨，京都府南部は曇り．京都府はすっきりしない天気ですね」と発話するが自然であり，北部／南部にそれぞれコメントを加えるのは冗長だということである．従って，コメント生成のためには情報の補完，および統合情報の信頼性を考慮したコメント付与対象の判定が必要となる．この補完技術に関して，4章で述べる．

以上のように，本稿ではコメント生成のための情報の信頼性の扱いを中心に論じ，最後に 5章でコメント生成ルールについて述べる．

3 信頼性と言語表現

3.1 情報の信頼性

我々は，情報の信頼性を可能性 (possibility) と確信度 (certainty) の 2 値で表現している [4]．人間がある情報源から情報を得たとき，その直接的な情報の可能性だけでなく，関連する情報の可能性についても主観に基づいて判断を下すことがよくある．例えば，「晴れ」という天気予報を得た場合，多くの人は「晴れ」という気象現象が生じる可能性が高いと考えるだけでなく，「曇り」の可能性もあると考えるし，また，「雨」の可能性はほとんどないとも考える．すなわち，「晴れ」という情報から「曇り」や「雨」といった関連する情報をも得ていることになる．そこで，取得した情報を (1) 式に示すような可能性分布 [3]

表 1: 可能性と叙述内容表現の対応

可能性	叙述内容表現
$0-a$	[コメント] ではない
$a-b$	[コメント] である可能性は低い
$b-c$	[コメント] である可能性は高い
$c-1$	[コメント] である

表 2: 確信度とムード表現の対応

確信度	ムード表現
$0-d$	[叙述内容] と思うのですが，自信がありません.
$d-e$	[叙述内容] かもしれません
$e-f$	[叙述内容] と思います
$f-1$	[叙述内容] でしょう

に展開し、取り扱うことにする[1].

$$\text{「晴れ」} \rightarrow \{1.0/\text{晴れ},\ 0.4/\text{曇り},\ 0.1/\text{雨}\} \tag{1}$$

このように，一地域に一般的に複数の天候とその可能性が存在することになる.

一方，確信度はその可能性の値がどの程度信頼できるかを示す値である. インターネット上のサイトから情報を得た場合，より多くのサイトから類似した情報を得た場合ほどその情報の確信度は高くなるものとする. つまり，サイト A, B, C からの情報が全て「晴れ」であった場合と，サイト C からの情報だけが「曇り」であった場合では，両者とも統合情報として「晴れ」を採用したとしても，後者は前者に比べて確信度が下がることとなる.

3.2 コメントの言語表現

次に，これらの信頼性とコメントの表現の対応を考える. 一般に文の意味内容は，ある事象を客観的に叙述する部分とその叙述に対する話者の心的態度を表す部分とに分けられる. 前者は叙述内容 (proposition), 後者は陳述 (modality) ないしムード (mood) と呼ばれる [5]. 特に，真とは断定できない叙述内容を述べる際のムードは「概言」と呼ばれ，その内容によって断定保留 (〜でしょう) や推定 (〜のようだ) などに分けられる [6]. そこで，可能性と確信度をそれぞれ叙述内容とムードに反映させる [4]. 定型的には，表 1, 2に示すような対応になる. ここで $0 < a < b < c < 1$ であり，$0 < d < e < f < 1$ である. しかしこれらはまだ不完全であり，一部不自然な表現となることは否めない. 実際に自然な表現とするためには，より詳細な対応付け [7] が必要となる. そのような柔軟な言語表現生成機構を確立することで，「真夏日になる可能性が高いでしょう」「傘を持っていったほうが安心だと思います」といった信頼性を反映した自然なコメントが生成できると考えている.

4 情報の補完

4.1 補完法

ユーザが要求している地域の気象情報がサイト上に存在しない場合を考える. 地域階層は木構造を想定しているので，上位は一意に決まる. 従って，もし上位地域に適切な情報があれば，その情報を以てその

[1] 可能性は確率とは違い主観的であり、「可能性が高い」は必ずしも「確率が高い」を意味しないが、それらの間には正相関があると言われている (possibility/probability consistency principle[3]).

まま下位を補完できる．つまり，京都市の情報が無い場合，代わりに京都府の情報を使う，ということである．

次に，下位地域から補完することを考える．問題を単純にするため，地域階層における下位地域への分割法は一意であるとする[2]．一般に上位地域に対して下位地域が複数存在し，上位地域の情報はそれぞれの下位地域の情報を統合したものとなるのが自然である．我々は [4] において，複数サイト間の情報を統合する手法を提案している．これは情報の信頼性を高めるための手法であったが，コメント生成という観点から再考察すると，この手法は情報補完のための地域間情報の統合にも応用できる．以下，情報統合による補完法を具体的に説明する．

補完に用いる下位地域の集合を r_j $(1 \leq j \leq J)$ とし，r_j での気象現象 $w \in W$ の可能性を $Pr_j(w)$，確信度を $Cr_j(w)$ とする．各 w について，上位地域の気象現象としての可能性を考えるが，統合により可能性が新しく生成されるのは不自然なので，w の可能性 $P(w)$ は $\{Pr_1(w), \ldots, Pr_J(w)\}$ のいずれかであるとみなし，その確信度を求めることとする．w の各可能性は既に下位地域の気象現象としての確信度を持っているので，それを上位地域での可能性の重みと考える．可能性 $P(w)$ が $x \in \{Pr_1(w), \ldots, Pr_J(w)\}$ である場合の確信度 $C(P(w) = x)$ は

$$C(P(w) = x) = \frac{\sum_{k=1}^{J} Cr_k(w) \cdot \mathrm{sim}\,(x, Pr_k(w))}{\sum_{k=1}^{J} Cr_k(w)} \tag{2}$$

で求められる．ここで関数 sim() は二つの可能性の一致度を示す関数であり，複数の地域で近い可能性を持っていればその確信度は高くなる．次はその関数の例である．

$$\mathrm{sim}(P_1, P_2) = e^{\gamma(|P_1 - P_2|)^2} \tag{3}$$

ここで $\gamma(< 0)$ は勾配を決める係数であり，γ が 0 に近付くにつれ，可能性の差に敏感になる．次に，$C(P(w) = x)$ を最大とする x を w の統合した可能性 $P(w)$ として採用し，その x に対応する確信度 $C(P(w) = x)$ を統合した可能性の確信度とする．さらに，補完に用いる下位地域数が多ければ得られた結果の確信度は高くなるので，確信度に「補完に用いた下位地域数 / 地域階層における全下位地域数」と正相関を持つ関数の値を掛け，得られた値を $P(w)$ の確信度 $C(w)$ とする．

以上により，上位地域の補完情報 $\{P(w), C(w) | w \in W\}$ が得られる．

4.2 コメント付与の妥当性

下位地域の情報を用いて上位地域の情報を補完した場合，一般には補完された上位地域の情報の信頼性は低い．システムはユーザに対して信頼性の高い情報を優先して提供するべきである．しかしその一方で，ユーザがシステムに上位地域の情報を求めている場合，情報の信頼性が低くとも，可能な限り上位地域の情報に対しコメントを提供することが望ましい．そこで，このトレードオフの下，システムがどの地域にコメントを付与するのが良いかを判定するために，各地域におけるコメント付与の妥当性 (adequacy) を計算する．

上位地域から見た全下位地域数 n が大きければ大きいほど，それぞれの下位地域 r におけるコメントの価値は下がると言える．そこでまず，情報検索や情報抽出の分野で一般的に用いられる F 値 [8] を応用し，下位地域 r の気象現象 w の総合的な信頼度 $Fr(w)$ を可能性 $Pr(w)$ と確信度 $Cr(w)$ から算出する [4].

$$Fr(w) = \frac{(\beta^2 + 1) \cdot Pr(w) \cdot Cr(w)}{\beta^2 \cdot Pr(w) + Cr(w)} \tag{4}$$

ここで $\beta(> 0)$ は可能性と確信度のどちらを重視するかを決定するパラメータであり，β が 1 のとき，調和平均を意味する．β を 1 より大きくすれば可能性を重視し，逆に β を 0 に近付けていくと確信度を重視することになる．

[2]実際は，一地域の分割は { 北部／南部 }, { 山間部／海沿い } のように複数種存在するが，その場合は補完の都度分割を固定すればよい．

表 3: コメント付与の妥当性 (1)

地域	気象情報	可能性	確信度	妥当性
京都府	晴れ	0.2	0.1	0.07
京都府	曇り	1.0	0.7	0.41
京都府	雨	1.0	0.8	0.44
京都府北部	雨	1.0	0.9	0.24
京都府南部	曇り	1.0	0.8	0.22

表 4: コメント付与の妥当性 (2)

地域	気象情報	可能性	確信度	妥当性
京都府	晴れ	1.0	0.1	0.09
京都府	曇り	0.5	0.3	0.19
京都府	雨	1.0	0.1	0.09
京都府北部	雨	1.0	0.9	0.24
京都府南部	晴れ	1.0	0.8	0.22

次に，r における w へのコメント付与の妥当性 $Ar(w)$ を，

$$Ar(w) = \delta_r(n) Fr(w) \tag{5}$$

で求める．ここで，$\delta_r(n)$ は分割数 n と負相関を持つ関数である．$\delta_r(n)$ として $1/n$ を用いた例を二つあげる (表 3, 4).

この妥当性が高い地域がコメントを付与するのに妥当な地域となる．この妥当性を指標とし，例えば京都府の情報を求めているユーザに対して，下位地域の情報が比較的類似している場合 (表 3)，「京都府北部は雨．京都府南部は曇り．京都府はすっきりしない天気です」というように上位地域にコメントをつけ，また，下位地域の情報が類似していない場合 (表 4)，「京都府北部は雨．京都府南部は晴れ．京都府北部では傘が必要です」というように下位地域にコメントをつける．このように適切に対象地域を判断することが可能となる．

また，情報に欠落がなく補完の必要がない場合でも，ユーザが要求している情報の信頼性が著しく低い場合には，この妥当性の判断手法を適用して上位または下位の情報を代替情報として提供することも考えている．

5 コメントの生成ルール

本稿では情報の信頼性に基づいたコメントの各種扱いを中心に論じてきたが，最後にコメントを生成するための知識について述べる．

我々は提示情報からコメントを生成するための知識，すなわち生成ルールをテキストから自動獲得することを検討している．一般に，インターネットから得られる気象情報は定型的であり，その獲得および分類は比較的容易だが，その定型情報 (ex. 最高気温 30 度) と日常的な事象 (ex. 薄着) を直接結びつける知識は気象情報のサイトにも一般のサイトにもあまり存在しない．そこで，直接結びつけるのではなく，段階を経ることを考える．具体的には，「最高気温 30 度」と「薄着」の間に「暑い」というような比較的気象に近い事象を挟む，ということである．つまり，我々は 2 つのレヴェルのコメントを考えている．例えば，「最高気温 30 度」という情報に対し，「暑い／真夏日」というのが第 1 のコメントであり，「薄着／汗が出る」というのが第 2 のコメントである．前者は気象に関する直接的なコメントで，比較的単純なルールにより生成できる．後者は気象現象から影響を受け得る日常的な事象に関するコメントであり，人間の持つ常識的なルールを必要とする．

表 5: コメント生成ルール 1

季節	天候	最高気温	…	コメント
春	–	$20 < x$	…	暑い
夏	–	$30 < x$	…	暑い
				…
春	晴れ	$15 < x < 20$	…	爽やか
				…

表 6: コメント生成ルール 2

事象	コメント
暑い	薄着をする, 汗が出る, …
爽やか	外出する, 気持ちがいい, …
雨	傘をさす, 水たまり, …
…	

第 1 のコメントを表す言葉は天気予報サイトの気象概況等によく現れるので，これらと気象情報の共起情報から生成ルールを自動獲得できると考えている．なお，これらは気象予報用語の定義的知識とも言えるので，現在は小規模なテーブル型ルール（表 5）を予報用語集 [9] から人手で作成して使用している．

一方，第 2 のコメントの生成には日常的な事象の推移を予測する知識が必要となる．特定ドメインに対する専門的解釈をコメントとするのであれば，既に研究の進んだエキスパート・システムのプロダクション・ルールと同じように質の良い知識源が存在するが，我々の目指す日常的なコメントのためにはもっと常識的な知識源が必要である．常識の構築の研究として Cyc [10] があるが，これはかなり大規模な人手での構築であり，コスト面からもあまり好ましくない．また，コメントの生成は Cyc のような深い知識でなくとも可能と考えている．そこで我々は，一般のテキストを知識源とし，テキスト上に直接現れる表層的な知識を数多く収集し，データベース化することで，日常的な事象推移を予測する研究をすすめている [11]．現在の実装では天気に関する知識を人手で構築したもの（表 6）を利用しているが，将来的にはこの技術によるデータベースの自動構築を検討している．

このように 2 種類のルールを使い分けることを前提とすれば，自動獲得したルールでのコメント生成が現実的なものになると考えている．

6 おわりに

本稿では，情報の直観的理解を支援するためにコメントを付与することを提案し，それに必要な，情報の信頼性に基づくコメントの表現法およびコメント付与の妥当性の判断法について述べた．これらは，情報の信頼性を可能性と確信度の 2 値で表現することを特徴とする．これにより，情報の信頼性に応じて適切にコメントを生成することができる．

人とコンピュータのインタフェイスの一つに音声対話がある．音声対話がインタフェイスとして優れているのは，それが人間対人間のコミュニケーションに近いからだと言えるが，単にインタフェイスが近いだけでなく，その内容も近づかなければ本当の意味での親しみやすいシステムとは言えない．そのためには人間の常識的な振舞いをエミュレートするような高度な知識処理が必要である．我々のコメント生成はまだまだ小規模なものであり，改善の余地が多々あるが，コンピュータとの人間らしいコミュニケーションへの第一歩としてその意義は大きいと考えている．

現在このコメント生成を，音声対話システム DUG-1[12, 13] に適用し，実際に天気予報案内システムの構築を行っている．今後は，コメント生成ルールの自動構築法の検討とともに，コメントがより自然な発話となるような言語モデルの検討，および対話実験による評価を行っていく予定である．

参考文献

[1] 佐藤理史:「ワールドワイドウェブを利用した住所探索」, 言語処理学会年次大会, pp. 447-450 (2000)

[2] Ashish, N. and Knoblock, C. A.: Semi-automatic Wrapper Generation for Internet Information Sources, in *Proc. CoopIS 97*, pp. 160-169 (1997)

[3] Zadeh, L. A.: Fuzzy Sets as a Basis for a Theory of Possibility, *Fuzzy Sets and Systems*, **1**(1), pp. 3-28 (1978).

[4] 松下光範, 佐藤浩史, 藤本和則, 加藤恒昭:「情報統合における情報の信頼性とその言語表現」, 第16回ファジィシステムシンポジウム予稿集 (2000)

[5] 寺村秀夫: 日本語のシンタクスと意味 I, くろしお出版 (1992)

[6] 益岡隆志, 田窪行則: 基礎日本語文法 — 改訂版 —, くろしお出版 (1992)

[7] Druzdzel, M. J.: Verbal Uncertainty Expressions: Literature Review, *Tech. Rep. CMU-EPP-1990-03-02* (1989).

[8] 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999)

[9] 気象庁 (監修): 予報作業指針 予報用語, 気象業務支援センター (1999)

[10] Lenat, D., Guha, R., Pittman, K., Pratt, D. and Shepherd, M.: "Cyc: Toward Programs with Common Sense," in *Communications of the ACM*, Vol. 33, No.8 (1990)

[11] 佐藤浩史, 笠原要, 松澤和光:「テキストからの表層的因果知識の獲得とその応用」, 信学技報 TL98-23 (1999)

[12] Nakano, M., Miyazaki, N., Hirasawa, J., Dohsaka, K. and Kawabata, T.: "Understanding Unsegmented User Utterances in Real-Time Spoken Dialogue Systems," in *Proc. ACL 99*, pp. 200-207 (1999)

[13] 堂坂浩二, 中野幹生, 宮崎昇, 安田宜仁, 相川清明:「制限知識下における効率的対話制御」, 言語処理学会第6回年次大会発表論文集 (2000)