ネットニュース情報の質問応答サービスシステム

2B－3

宮部 隆夫
NEC 関西 C&C 研究所

1 はじめに

近年、ネットワークを介して膨大かつ多種多様な情報が流通している。そこから有効な情報を取得するためには、情報を整理選別するための専用機能が、必要になって来た。ネットニュースの記事のように、情報が散在し、書式が不正確なテキストに対しては、検索収集するだけでなく内容の圧縮や選択も不可欠である。そのためにニュースダイジェストを自動作成するシステム [佐藤 等 95] [佐藤 95] が提案されている。これは特定情報 (インターネットアドレスや製品仕様など) の抽出には有効であるが、対象記事や処理機能を限定しているために、情報洩れも多く拡張が難しい。

本稿では、ネットニュース記事の主題とその関連情報を整理し、利用者からの関連する質問に整理情報を提示するシステムを提案する。特に、複数の記事から主題情報とその関連情報を抽出し、両者を関係付けて提示する情報整理方式を説明する。

2 ネットニュース記事の特徴

ネットニュース記事は、電子メールと同様に不特定多数者間の対話型のテキストであり、書籍や雑誌、新聞などの文書とは異なる特徴を有する。

多様性 (個人差)： 不特定多数の人間が、明確な基準を持たずに各自の趣向に従ってテキストを作成する場合が多い。そのため、特殊な専門用語が含まれたり、記号の使い方や記述スタイルなどが記事により異なる。

品質 (冗長性、非文、口語)：推敲が不十分なため、情報が未整理で、文法的に不適切な低品質のテキストが多い。例えば、主題が簡潔に明言されていない、具体的対象が記載されていない、助詞や述語などが過度に省略されている、記述ミスを含む、などの問題点を含んでいる。

主題の並列・継続性：1記事内で複数の主題を提示している場合や記事中では話しが閉じていない場合が多い。複数の記事同士が密接に関係しており、1記事内だけでは情報が不足しており、意味が曖昧になる。

3 基本方針

前章で述べたニュース記事内の特徴を考慮し、適切な情報を抽出するために、**表層表現中心**、**主題の範囲**の抽出、**記事間の相互制約**の3つの基本方針に従った処理方式を提案する。以下に内容とその根拠とを述べる。

1. **表層表現中心**：ノイズに強く、また、情報を付加しやすい表層表現 (文字列) 中心の処理を行なう。前章でも記述したように、ニューステキストは多様性に富んでいる。システムにとっては、多くの未知語や多様な表現への対応が不可欠である。また品質の点では、文法誤りや記述ミスを含む文章が予想されるため、表層表現重視の処理にする。
2. **主題の範囲**：処理結果に曖昧性を残し、不確実な部分も含めて利用するために、**主題の範囲**を抽出することにする。明白に不適切と判定された部位を削除するとともに、適切な表現を含む範囲を収集して判定選択する。ニューステキストは、多様性に富み、システムにとって未知な表現 (単語) や不明確な記述も多く、また、品質が低いので、確からしい箇所を網羅する方が現実的である。
3. **記事間の相互制約**： 参照している (引用関係にある) 複数の記事を結び付け、双方の記事からの制約により主題範囲を推定し、記事間の関係づけを行なう。ニュース記事が対話型の情報表現であり、複数の記事が複数の主題について相互に関係している。この主題の並列・継続性の問題に対処するには、関連する記事を同時に処理する必要が生じる。

4 処理方式

ニュース記事を整理し利用者からの質問に応じてその内容を回答するシステムのアーキテクチャーを、図1及び図2に示す。

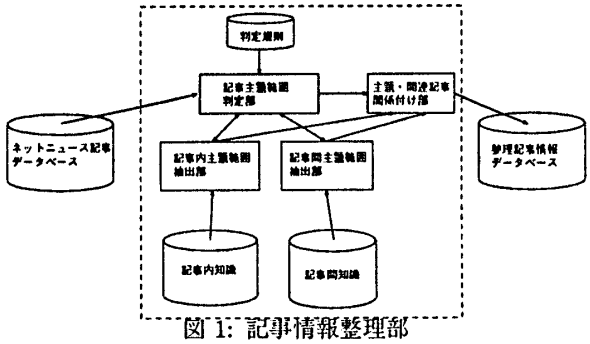

図 1: 記事情報整理部

図1の**記事情報整理部**では、ニュース記事のデータベースから順次記事を呼びだし、記事内容の分割や要約、他の記事との関係づけを行なって情報を整理し、**整理記事情報**を作成する。整理は記事の主題情報を中核にして進む。記事内及び記事間知識から各主題範囲抽出部が主題範囲の候補を抽出し、判定部が規則を利用してその可否を判定する。必要に応じて別候補探索も指定する。適切な主題範囲に対しては、関係づけ部が記事内間の知識を元に、関係を指定し、結果を整理記事情報として格納保存する。

図2の**質問応答部**では、上記の記事情報整理部で作成した整理記事情報を、利用者からの質問に応じて回答する。

A Question and Answering System for Network-news Information

Takao MIYABE
Kansai C&C Research Labs., NEC Corp.

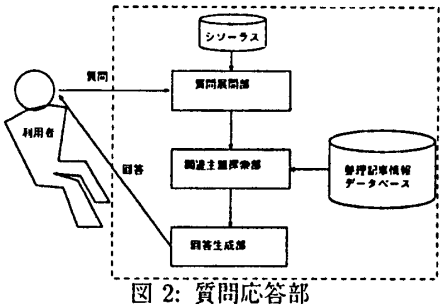

図 2: 質問応答部

5 利用知識

ここでは、記事情報整理部が主題抽出に利用する知識を示す。

5.1 記事内知識

記事内の知識を表1に示す。知識は各種カテゴリ (とサブカテゴリ) 毎に階層表現され、表現分類の下に各種言語表現 (文字列) が記載される。**記事種類**は記事の種類を定める知識であり、記事間の関係づけに用いる。通知や質問などのキーワードが記載される。**記事構造**は、記事内のテキスト構造を推定し、主題範囲の分割や不用情報の削除に利用する。枠構造には挨拶などに関する規約や表現が、箇条書や並列には項や助詞の表現が、補足には例示を示す関係表現が記載される。**主題述部**は、主題に対応する述部の表現が記載され、主題範囲推定に用いられる。情報の提示や疑問や判定などの表現を含む。

表 1: 記事内知識

カテゴリ	サブカテゴリ	表現分類 (表現)
記事種類	通知, 質問,..	会議案内, 質問,....
記事構造	枠構造, 箇条書 並列, 補足,..	挨拶 (はじめまして..), 項番 (数)), 項記号 (·), 並列助詞, 例示,..
主題述部	提示, 疑問, 判定,..	教示 (教えて), 疑問符 (?), 真偽,..

この知識の利用例を表2に示す。**質問**という文字列により記事種類が、また挨拶「はじめまして」及び**項番**「1),2)」情報から2つの段落からなる記事構造が推定できる。更に**提示**,**疑問**に関する表現「教えて」「?」から、主題範囲として「XXX」「yyy」とが求まる。

表 2: 記事内知識利用例

はじめまして,..... 以下の質問について、どなたか.... 1) XXX について教えてください。..... 2) yyy はどうですか ? .....

5.2 記事間知識

記事間の相互制約の知識を、表3に示す。形式は前述の記事内知識と同様である。**引用知識**には、記事及び範囲の指定知識がある。(in-reply-to) などの記事指定知識や、引用範囲を特定するための引用記号表現が記載される。**記事間構造知識**は、引用の構造を推定するための知識である。記事の部分毎に別箇所から引用される場合 (部分) や、複数記事から引用される場合 (重なり)、再引用 (引用の引用) される場合 (多重) を検出し、それぞれを主題範囲の分割、縮小、共有とする。

表 3: 記事間知識

カテゴリ	サブカテゴリ	表現分類 (表現)
引用	記事, 範囲	定義, 表記,...
	範囲指定	記号 (>), 識別子,..
記事間構造	並列, 多重,..	部分, 重なり, 多重,..

この知識の利用例を表4に示す。上段の記事1に対して、下段の記事2中の (in-reply-to 記事1) により、記事間の引用が分かり、さらに>という引用記号と文字列一致から、引用範囲が求まる。結果として、記事2が記事1中の主題 XXX に関する回答として関係付けられる。

表 4: 記事内知識利用例

(記事1) XXX について教えて下さい。,..... .... ところで yyy はどうでしょうか ?
(記事2 in-reply-to 記事1) > XXX について教えてください。..... XXX とは .....

5.3 判定規則

表5に、主題範囲候補の判定のための判定規則を示す。主題範囲内において、条件内の表現を探索し、該当処理を行なう規則となる。具体的対象が存在する場合は、適切とされる。他方、指示表現の場合は不可として、範囲を拡大して別候補を探索する。また、不要語のみの場合は、不適切とされ、別候補を探索する。

表 5: 判定規則

条件		判断・処理	
カテゴリ	代表表現	判断	処理
具体的対象		可	主題範囲
指示表現	指示詞, 参照,..	不可	範囲拡大, 別候補探索
不要語	述語, 抽象表現,..	不可	別候補探索

6 おわりに

本稿では、ネットニュース情報を質問応答を介して提示するシステムについて述べた。主題情報とその関連情報とを抽出して関係付けをする情報整理方式を中心に、システムの処理方式と利用する知識とについて説明した。

参考文献

[佐藤 等95] 佐藤, 佐藤, 篠田「電子ニュースのダイジェスト自動生成」 情処学会論文誌 Vol.36,No.10, pp.2371-2379

[佐藤95] 佐藤「ネットニュースダイジェストの自動生成」自然言語処理の応用に関するシンポジウム95,pp.81-88