5B－7　自然言語による協調作業支援データベースシステムのためのデータモデル及びデータベース自動生成システムの設計

岡本 東　樋地正浩　佐藤 究　宮崎正俊

東北大学大学院情報科学研究科

1　はじめに

協調作業を行っていて作業内容に関する問題が発生した場合、他の作業者に質問し回答を得て問題に対処するといったことはしばしば行われる。ここで、過去にあった問題とその対処の方法が蓄積されていてそれを参照することができれば、迅速に問題解決を行い作業を効率的に進めることができる可能性がある。

このようなデータベースを用意しようという試みは一般によく行われているが、しかしながら、データベースシステムだけを用意しても、データを蓄積するためのよい手法がなく、手でデータを入力してデータベースを作成しているというのが現状である。

我々は協調作業支援システムの一部として過去にあった問題と対処法に関する情報を蓄積するデータベースシステムの開発を行っており、ここで、先に述べた問題点を解決するために、人間同士でやり取りされた情報の中からデータを抽出し蓄積するための研究を進めている。本稿では、データの抽出方法及び格納型式について検討し、それに基づく試作システムによる実験について述べる。

2　協調作業支援データベースシステム

本研究では以下のような特徴を持つ協調作業支援データベースの開発を目標としている。

- 自然言語でやりとりされる情報からデータを収集し蓄積する。
- 自然言語でやりとりされる質問を蓄積されたデータと照合し、適切な解答があればそれを返す。

このようなデータベースシステムを作成する上で、以下の点について検討しなければならない。

- どの様にして、自然言語でやりとりされている情報を収集するか。
- 情報を再利用するために、どのような形で保存するか。

今回は自然言語でやりとりされる情報として netnews や mailing-list の記事から要旨を抽出し解析するシステムを試作し、これを用いて実験を行って、上記の点についての検討を行った。

3　データモデル

netnews の記事を解析する研究としては、参考文献[1]などがあり、形態素解析を行わずに文字列パターンからダイジェストやサマリを抽出することに成功している。

しかし、本研究では、抽出した情報を再利用する必要があるため、上記のようなパターンマッチの手法を利用して抽出するだけでは不十分であり、それを再利用可能な形に構成しなおさなければならない。

具体的には、次のような手順で行われる。

最初に、パターンマッチにより記事全体の中から、質問の記事ならば質問文、回答の記事ならば回答文を抽出する。ただし、質問の記事の場合、質問文のみでは情報として不十分であり、その質問をするに至った状況も必要となることが多い。よって状況説明の文もあわせて抽出する。

Data model and automatic building databases for collaboration system

Azuma Okamoto, Masahiro Hiji,
Kiwamu Sato, Masatoshi Miyazaki
Graduate School of Information Sciences,
Tohoku University

次に、抽出された文を再構成し再利用可能な形にするが、再構成は、抽出した文同士の関係のリンクを作成することと、抽出した文中の各要素 (単語) とそれらの関係の構造を作成することの二つに大きく分けられる。前者は前段階のパターンマッチの段階で決定され、後者は形態素解析を行って決定する。

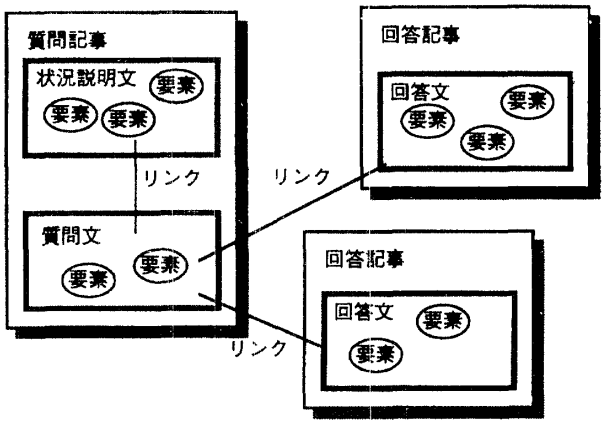

図 1: データモデル

どのような要素が必要となるかはデータベースシステムで扱う内容によって異なってくる。計算機管理に関する内容を扱った場合の状況説明文の要素の例を以下に示す。

表 1: 状況の要素の例

ハード	NWS-830, NWS-4000, NWS-5000, ...
OS	NEWS-OS 4.2.1, NEWS-OS 6.1, ...
アプリ	iij-ppp, Netscape Navigator, ...
その他	移植, インストール, デバッグ, ...

4 データベース自動生成

前節で述べたデータモデルに従ったデータベースを自動生成する実験を行った。

自然言語文としては、netnews の技術系の newsgroup の記事を用いた。単一の newsgroup の記事を元に、状況説明文・質問文・回答文の特徴的なパターンと、それらの文を形態素解析するために必要な文法を作成し、それらを用いて同じ newsgroup の別の記事に対して処理を行った。形態素解析については、既出の単語のみを用いた文でないと解析ができずデータベースの自動生成という目的を達することができないため、文法に適合しない部分は未知語 (名詞) として扱い、新しい語として登録する。ただし、未知語が長くなってしまう場合や句読点まで到達してしまう場合には、未知語として扱うのをやめてその文の解析を打ち切る。

5 実験結果

パターンマッチによって、状況説明や質問の文の抽出は可能であったが、質問と対応する回答とのリンクが困難であった。これは、質問の記事に対するフォローの記事であっても、必ずしも回答が書かれているわけではないのが主な原因である。

また、形態素解析を用いた再構成の処理は、定形パターンの文章の解析となるために比較的単純な規則で十分な解析を行うことができた。ただし、未知語を次々と蓄積していったため、同じ意味の違う語や少しずつ違う表現や誤字などが蓄積されていくという問題が発生した。これらは人間が分類してやらない限り、検索に利用するのは困難である。

6 おわりに

協調作業支援システムの一部である協調作業支援データベースのデータモデルとデータベースの自動生成に関して述べた。

実験によって、状況説明や質問をデータベース化することができることは確認できた。また、今回は質問と回答がきちんと対応するリンクを作成するまでには至らなかったが、netnews ならば、ヘッダ情報などの自然言語文以外の手掛りを用いて関連事項のリンクを作成するのは比較的容易であり、実際に協調作業支援システムで用いる場合にも同様な情報を付加する仕組みを取り入れれば実用になるレベルに達すると考えられる。

参考文献

[1] 佐藤円, 佐藤理史, 篠田陽一:
電子ニュースのダイジェスト自動生成, 情報処理学会論文誌, Vol.36, No.10, pp2371-2379, 1995.