電子メールの文書からの関係情報の自動抽出

4Q－6

塚本雄之　河合敦夫　天野則幸　山内文博　山下禅　椎野努
三重大学　工学部

1　はじめに

大量の文書データの中から、あらかじめ目的とした情報のみを取り出してくる技術、すなわち内容抽出の技術は、文書情報の整理やデータベースの自動的な構築のためには、必須の技術である。こうしたテキストからの内容抽出の研究は、学術論文[1)]、特許[2)]、製品の紹介記事[3)]等で研究されている。しかし、これらは、いずれも、印刷出版物の文からの内容抽出における研究である。これらの文書と、電子メールからの文書を比較すると、大きく２つの点で異なる。

１つは、前者が、いわゆる主語、述語等を備えた文のみからの内容抽出であるのに対し、電子メールでは、表、箇条書き、単語の羅列といった文以外の構成要素を考慮する必要がある点である。もう１つは、前者が、印刷出版物であるため、著者、校正者、査読者等の推敲によって、文書の表現、使用する単語の使い方等がチェックされている。これに対して、電子メールでは、こうした多段階による推敲過程は、通常存在しない。ここでは、特に前者に着目した内容抽出方式について発表する。

2　電子メールの文書構造と処理手順

電子メールは、日本語から記述されている。しかし、いわゆる、主語、述語等からなる文のみから構成されているわけではない。我々は、電子メール中に出現する文書構造を以下の３つに分類した。

①表部分：文字等の物理的（２次元的）配置が意味を持つ。文および単語の羅列からなる。この中には、いわゆる箇条書きも含む。

②文章：文字等の物理的配置が意味を持たない。文のみから構成される。

③その他：（①～②以外の部分、名詞の羅列等）

また、通常の電子メールでは、すべての文書はテキストファイルの形式、すなわち文字コードから構成されるデータで記述されている。言い替えれば、表（や図）は、一太郎や花子、オアシス等のワープロ独自のフォーマットで記述されているわけではなく、図１に示すように単語や文字の物理的配置によって表現されている。以上の観点にたって、以下に示す処理の流れを採用している。

①電子メール文書の形態素解析を行う。

②表部分全体を行単位で認識する。

〈電子ﾒｰﾙの文書例〉

```
27  FBH01433  8/03  34 RA21他売ります。
    塚本　雄之
◎PC9801RA21本体
  11万円で。
◎120ﾒｶﾞHDD
  緑電子製、SCSI NOVA V120
  新品同様、購入後　1か月。
  多少のｱﾌﾟﾘｹｰｼｮﾝｲﾝｽﾄｰﾙ済み。
  4万3千円で。
◎3.5ｲﾝﾁFDD　緑電子製、1FDD New Little F
  1万2千円で。
◎内部増設4MRAM ﾒﾙｺ製 EDA4000
  1万円で。
  いずれも、箱、ﾏﾆｭｱﾙ完備。
取りにこれる方には、ｹﾞｰﾑｿﾌﾄ（ちょっと古いけど）さしあげます。その他は、宅急便で送料無料で、送ります。
  FBH01433 TSUKA
```

〈内容抽出結果〉

名　称	型　番	ﾒｰｶｰ名	価　格
ﾊﾟｿｺﾝ本体	PC9801RA21	ＮＥＣ	110,000円
ﾊｰﾄﾞﾃﾞｨｽｸ	NOVA V-120	緑電子	43,000円
ﾌﾛｯﾋﾟｰﾃﾞｨｽｸ	New Little F	緑電子	12,000円
増設ﾒﾓﾘ	EDA 4000	メルコ	10,000円

図１　電子メールの文書例と内容抽出結果

Information Extraction from Electronic Mail
Takeyuki Tsukamoto, Atsuo Kawai, Noriyuki Amano, Fumihiro Yamauchi, Yuzuru Yamashita, Tsutomu Shiino
Faculty of Engineering, Mie University
1515, Kamihama-cho, Tsu 514 Japan

③表部分を、まとまりのある行ごとに分割する。この分割を、ここではブロックと呼ぶ。
④表部分中の項目間で、縦の列に関連があれば、それを認識する（図２は、②～④を図式的に表現している。図２のプロセスは、物理的レイアウトから、論理的な関係を復元しているとも考えることができる）。
⑤表部分からの内容抽出を行う。
⑥文章からの内容抽出を行う。

②～④は、文字、品詞、および単語の持つ意味属性の物理的配置の特徴を利用して認識を行っている。例えば、「特殊文字（◎）が、同一列に並ぶ現象を利用して箇条書きを認識する（図１参照）」、「縦の同一列にメーカー名が並ぶことによる表構造の認識」などのルールを考えることができる。

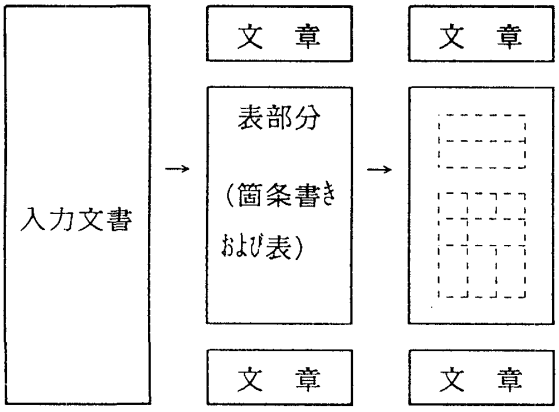

図２　文書の分割

3　処理対象文書

電子メールには、さまざまな文書が存在する。このうち、記述対象がある程度限られた範囲に限定できる、記述内容に一般性がある（個人間の私信ではない）、文書が公開され実験対象としてまとまった数が入手できる、といった観点から、パソコン通信の中古品売買情報（売ります買います）に着目した。図１に電子メールの文書と内容抽出結果（例）を示す。ここでは、特に、中古品の中でも、パソコンを取り扱っている文書に限定し、内容抽出の目標を、パソコンの本体や付属品の名称、型番、メーカー名、価格とした。

4　抽出の手がかり

内容を抽出するための手がかりとしては、
①単語の字面そのもの
②単語に付与した意味属性
③単語を構成する文字列の種別から推測
　例：電子メールのユーザーＩＤ（３桁の英文字とそれに続く５桁の数字）
④表部分の認識結果の利用
　同一ブロック内にある抽出情報の関連づけ
　項目名（表のｈｅａｄｅｒ）の利用
　表中の同一列の単語の意味属性からの未知語の推測
⑤手がかりとなる単語（または意味属性）が、前方または後方に出現(例：主記憶 640k、走行距離は30,000kです、値段:780k)
⑥文の構文パターン
⑦対象分野についての知識を使った推論規則

を考えることができる。

内容抽出の処理は、情報の抽出（いわゆる文書からのキーワード抽出）と、抽出された情報の関係付け（いわゆる構造化キーワードの構成）の２つに分けて考えることができる。こういった点から考えると、前者については、①～④、後者については、④～⑥が関連する。

5　おわりに

現在、多くの電子メール文書に適用して評価を行うとともに、処理系の開発を行っている。

参考文献

1)猪瀬博、斉藤忠夫、堀浩一：シナリオを用いる論文抄録理解・作成援助システム、情報処理論文誌、Vol.24,No.1,pp.22-29(1983)
2)高松忍、日下浩次、西田富士夫：技術抄録文からの関係情報の自動抽出、情報処理論文誌、Vol.25,No.2,pp.216-224(1984)
3)松尾比呂志：抽出パターンの階層的照合に基づく内容抽出法、情処ＮＬ研究会、Vol.99,No.2,(1994)