電子ニュースのダイジェスト自動生成

佐藤 円[†] 佐藤理史[†] 篠田陽一[†]

現在，電子ニュースを通じて多くの情報が流通し，多くの人々がその情報を利用している．この電子ニュースは，新しいマスメディアであり，従来のテキスト情報マスメディアにはない，優れた特徴を持っている．しかしながら，現在のニュースリーダは，その特徴や利用者の要求に合致した，適切な機能を提供しておらず，読者にとっては，必ずしも利用しやすい情報メディアとはなっていない．我々は，電子ニュースを利用しやすい情報メディアにするためには，そのダイジェストを提供することが不可欠であると考える．ダイジェストとは，元になる情報をコンパクトにまとめ編集したものであり，情報全体の俯瞰やエッセンスの把握，情報の取捨選択の際に，優秀なナビゲータとして機能する．本研究では，電子ニュースに対して，このようなダイジェストを自動生成することを提案し，その一つのプロトタイプとして，会告記事用ニュースグループ fj.meetings のダイジェストを自動生成する方法を示す．ダイジェストの自動生成を実現する中心的な技術は，サマリーの自動抽出技術であり，会告記事にみられるスタイル上の特徴，言語表現パターンを利用することにより，実用に十分な精度でサマリーを抽出できることを示す．本方式で自動生成されたダイジェストは，WWW のクライアントプログラムで読むことができる．

Automatic Digesting of the NetNews

MADOKA SATO,[†] SATOSHI SATO[†] and YOICHI SHINODA[†]

The amount of information in the NetNews is already enormous, and more and more people are using the NetNews as the mass media. However, the NetNews system does not provide efficient means to support users enough. We propose the digest of the NetNews as the efficient means to navigate users. A digest is an edited collection of summaries of original texts. Digests help users when they overlook the NetNews as a whole, grasp the essence of information, and choose articles from many candidates. Automatic generation of digests is possible for the NetNews because the NetNews exist as electronic text. This automatic generation saves a lot of time and costs compared with manual generation. Another important feature is that the automatic generation of digests is the automatic information reproduction. In this paper, we described the design and implementation of the digest system for a conference announcement newsgroup, fj.meetings. The key technology of the digest system is the extraction of prespecified information (summary) using article style information and expression patterns that are distinctive to articles in the conference announcement newsgroups. The results of summary extraction evaluations demonstrate high recall and precision. Edited summaries can be read by WWW client programs (e.g. xmosaic).

1. はじめに

現在，電子ニュースを通じて多くの情報が流通し，多くの人々がその情報を利用している．例えば，電子ニュース fj の 1 カ月の投稿量は，約 17000 件，約 36MB であり，これは，ほぼ新聞の全国紙 1 紙の 1 カ月分に相当する分量である[1)]．そこに掲載される情報内容も多様で，記事のタイプも，情報告知型，質問-応答型，議論・おしゃべり型などが混在している[1)]．これらのことから，電子ニュースでは，量的にも質的にも，すでにマスコミュニケーションが行われていると言うことができるだろう．

マスメディアとしての電子ニュースは，新聞や雑誌といった他のテキスト情報マスメディアと以下の二つの点で大きく異なっている．

(1) 情報が，電子メディア上で表現されている．このことより，
  (a) 情報をコンピュータ・ネットワークを介して伝搬することができるので，広い地域に速く伝えることができる．
  (b) 情報を読むために，ニュースリーダと呼

† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology

ばれるプログラムを必要とする．

(2) 情報の直接発信の機会を利用者全員に開いている．すなわち，電子ニュースでは，情報（記事）の取捨選択が行われず，かつ，発信者が書いたそのままの状態で読者のもとに配送される☆．

このように，電子ニュースは，従来のテキスト情報マスメディアにはない新しい特徴を持っている．このため，この電子ニュースを通じて，今後，新しいマスコミュニケーションの形態が実現される可能性があると考えられる．しかし，現状では，電子ニュースの特徴は十分に活かされているとはいえない．その一つの大きな原因は，現在のニュースリーダが，電子ニュースの特徴と利用者の要求に合致した，適切な機能を提供していないことにあると考えられる．

現在のニュースリーダの提供している機能は，基本的に，ニュースグループを選択する機能とニュースグループ内の記事を到着順に表示する機能の二つである．これらの機能により，読者はすべての記事にアクセスすることは原理的には可能であるが，例えば，ニュース全体で今何が起こっているのかを知ること（全体の俯瞰）や，求める情報に素早くアクセスすることは，それほど容易ではない☆☆．このため，現在の電子ニュースは，読者にとって，必ずしも「使いやすい情報メディア」とはなっていないのである．

従来のテキスト情報マスメディアでは，上記の問題はどのように解決されているのであろうか．多くのマスメディアには，元になる情報をコンパクトにまとめ編集した「ダイジェスト」が付加されている．例えば，雑誌や書籍の目次は，典型的なダイジェストである．これはテキスト情報マスメディアに限られたものではない．例えば，テレビに対するテレビ番組表，テレビニュースにおけるヘッドラインなども，ダイジェストである．このようなダイジェストは，情報全体の俯瞰やエッセンスの把握，情報の取捨選択の際に，優秀なナビゲータとして機能する．特に，読者が，すべての情報ではなく，限られた部分だけを必要としている場合，その必要部分を見つけ，それに到達することを支援する際に，絶大な威力を発揮する．

我々は，電子ニュースを「利用しやすい情報メディア」にするためには，このダイジェストが不可欠であると考える．なぜならば，電子ニュースの利用においては，記事を網羅的に読むのではなく，読者が自分に必要な記事を拾い読みするといった取捨選択的読み方が一般的であるからである．しかしながら，電子ニュースのダイジェストを人手で作るということは，作成に時間がかかり，新たな費用を必要とするという点から，現実的ではない．

そこで，我々は，電子ニュースのダイジェストを自動生成することを提案する．電子ニュースは，既存のテキスト情報と違い，情報がはじめからオンラインテキストとして存在するため，原理的には，ダイジェストの完全自動生成が可能である．また，これが実現できた場合は，以下のようなメリットが得られる．

(1) 大量の情報に対するダイジェスト作成・更新が，短時間でできるため，電子ニュースの特徴の一つである迅速な情報流通が，ダイジェスト作成作業に費やす時間のために妨げられることがない．

(2) ダイジェスト作成コストがほとんどかからないため，作成作業に伴う費用を賄うための課金の必要がない．

また，このダイジェストをハイパーテキストとして作成すれば，ダイジェストから簡単にオリジナル記事に移動することが可能となる．

本論文では，このような電子ニュースダイジェストの一つのプロトタイプとして，会告記事用ニュースグループ fj.meetings のダイジェストを考え，これを自動生成する方法について述べる．

2. fj.meetings ダイジェスト自動生成システムの概要

2.1 fj.meetings とそのダイジェスト

ニュースグループ fj.meetings は，会告記事用のニュースグループであり，そこには，主に，会議告知記事と論文募集記事（以下では，この両者をまとめて会告記事と呼ぶ）が投稿される．その多くは，日本語で書かれたものであるが，英語で書かれたものも混在する．**図1** に，日本語会告記事の典型例を示す☆☆☆．

このような会告記事は，その会議に参加しようと考えている人々とっては重要な情報を提供するが，それ以外の人々にとっては，全く不用であり，読む必要がない記事である．そのため，読者は，自分の興味のあるものだけを拾い読みし，あとは読み飛ばしたいという希望を持っている．

☆ パソコン通信の場合，システムオペレータにより，特定記事の排除やニュースグループの運営方針が決められることがあるが，他の文字メディアの状況と比較すると，電子ニュースの情報は，「編集」されていない「未加工情報」であるということができる．

☆☆ いくつかのニュースリーダでは，ニュースグループ内の記事をサブジェクトごとに整理するなどの機能を提供しているが，利用者の要求を満たす十分な機能を提供しているとはいえない．

☆☆☆ HASIDA.94Jun25223514@etlcom.etl.go.jp

第4回マルチエージェントと協調計算ワークショップ (MACC'94)

** 発表・参加募集のご案内 **

日本ソフトウェア科学会「マルチエージェントと協調計算研究会」(略称 MACC) では、下記要領で第4回ワークショップを開催します。どうぞ奮ってご参加下さい。

1。日時　1994年10月11日(火)、12日(水)、13日(木) (2泊3日)

2。会場

名称: ラフォーレ那須
所在地: 栃木県 那須郡 那須町 湯本 206-959
(中略)

5。論文発表

・論文発表希望者は、A4版2ページ程度のアブストラクトを6部、8月25日まで にプログラム委員長に送付して下さい。
(以下略)

図1　記事の例
Fig. 1　Typical article in Japanese.

このような記事の取捨選択には，記事の本文をすべて通読する必要はなく，限られた情報だけから判断できることが多い．すなわち,「どのような会議がいつどこで開催されるか」という情報（会議のサマリー情報）がわかれば，多くの場合，読者は，その会議が自分にとって興味のありそうな会議であるか否かを判断できる．これは,「各記事から，会議のサマリー情報を抜き出し，それをリストアップしたダイジェストを作成すれば，それは，読者による記事の取捨選択を効果的に支援することができる」ということを意味している．

2.2 ダイジェスト自動生成システムの概要

上記の考え方に基づき，fj.meetingsのダイジェスト自動生成システムを作成した．その概要を，**図2**に示す．本システムは，以下の三つのモジュールから成る．

(1) **記述言語の判定** — 日本語記事か英語記事かの判定を行う☆．
(2) **サマリー抽出** — 上記結果に従い，各記事からその記事（会議）のサマリー情報をその言語特有の方式で抽出する．
(3) **ダイジェスト編集** — 各記事から抽出したサマリー情報を編集し，ダイジェストを作成する．

以下の節では，サマリー抽出とダイジェスト編集について詳しく述べる．

☆ 記事に含まれる日本語文字のバイト数とASCII文字のバイト数を数え，後者が前者の3倍を越えた場合は，英語記事と判定し，それ以外の場合は日本語記事と判定する．

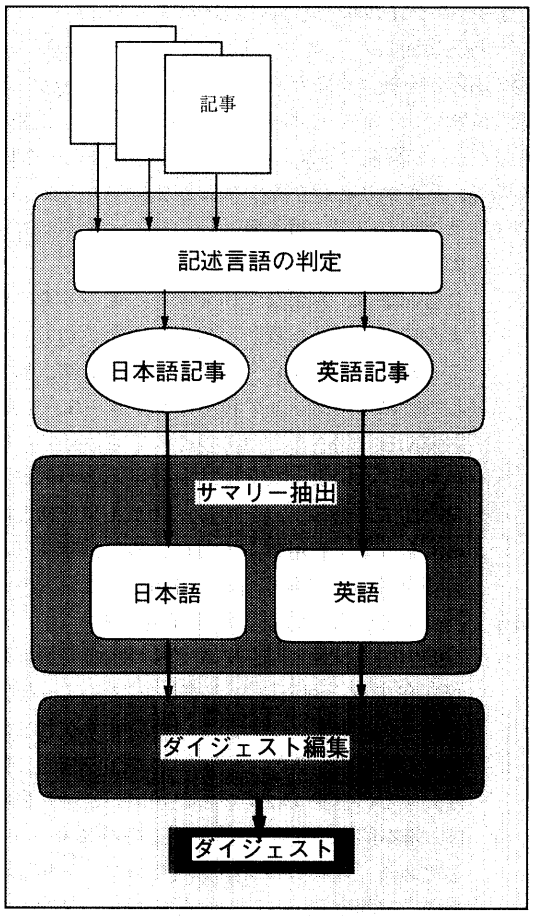


図2　ダイジェスト機構の構成図
Fig. 2　Digesting system.

なお，本システムは，現在，JAISTにおいて試験運用しており，WWWを通じて，自動生成されたダイジェストにアクセスすることができる☆☆．

3. サマリー抽出

サマリー抽出では，各記事から，その記事（会議）のサマリー情報として，以下の情報を抽出する．

(1) 記事種別（会議告知記事または論文募集記事）
(2) タイトル
(3) 開催期日
(4) 開催場所
(5) 論文募集締切期日（論文募集記事のみ）

サマリー抽出の方式は，日本語記事用，英語記事用でそれぞれ異なっている．以下では，紙面の都合上，日本語記事からのサマリー抽出についてのみ述べる．

☆☆ http://www.jaist.ac.jp/user/sato/nnad/digest-j.html

3.1 スタイル情報と言語表現パターン

会告記事からのサマリー抽出はどのような方法で可能であるかを検討するために，まず，fj.meetingsに実際に流れた114件の日本語会告記事を収集し，調査を行った．この調査の結果，以下の二つが，サマリー抽出の際に，手がかりとなりそうなことが明らかになった．

(1) センタリング，箇条書といった，テキスト表示のスタイル．
会告記事は，重要な情報を読者が見つけやすいように，テキスト表示上の工夫がなされていることが多い．例えば，タイトルは，通常，独立した行にセンタリングか左寄せで書かれ，また，開催期日・場所などは，多くの場合箇条書で書かれる．このようなテキストの表示上のスタイルに関する情報（以下では，スタイル情報と呼ぶ）をうまく利用することによって，サマリー情報を抽出すべき場所をかなり限定できる．

(2) 抽出する情報に特有な言語表現パターン．
抽出すべき情報には，それぞれ特有な言語表現が存在する．例えば，タイトルには，『～研究会』，『～シンポジウム』といった言葉が含まれることが多く，また，その後には『ご案内』，『開催』などの言葉が続くことが多い．日付の表記は，基本的に，『X年Y月Z日』のバリエーションである．これらの言語表現をパターン化したもの（言語表現パターン）とのパターンマッチングによって，多くの場合，抽出すべき情報を特定することができる．

このような観察に基づき，サマリー抽出の際に利用するスタイル情報と言語表現パターンを整理した．その概略を**表1**に，タイトルの言語表現パターンの概略を**表2**に示す．

3.2 サマリー抽出アルゴリズム

サマリー情報の抽出アルゴリズムの概要を以下に示す．

(1) **行スタイル情報の判定**
各行が，1）センタリング行，2）左寄せ行（ノーマル），3）右寄せ行，4）タブ行☆，5）境界線行☆☆，6）空行，のうち，どれに相当するかを調べる．判明した行スタイルを，行ラベルとして各行に付加する．

(2) **記事種別の抽出**
記事全体をスキャンし，会議開催告知パターンと論文募集パターンの存在の有無を調べ，記事種別を判定する．

(3) **タイトルの抽出**
記事の先頭行から，箇条書の手前までを順に調べて，タイトルの抽出を試みる．
(a) その行の行ラベルが，センタリング行・タブ行・左寄せ行以外であれば，スキップする．
(b) その行が，通知パターンか会議種類パターンだけであれば，その直前の空行でない行をタイトルとして抽出する．
(c) その行が，タイトルパターンを含めば，その行をタイトルとして抽出する．なお，その行の直前の行に，タイトル先端パターンがあれば，それをタイトルの先端に付加する．
(d) その行が，記号飾りのあるセンタリング行であれば，記号飾りを除いたその行全体をタイトルとして抽出する．

(4) **箇条書部分からの情報抽出**
ステップ3で箇条書に到達した場合，それ以降の箇条書部分から，以下の方法で，開催期日・開催地・論文締切・タイトルを抽出する．
(a) 箇条書のラベルより，その箇条から抽出すべき情報を判定する．
(b) 開催期日を抽出する場合は，ラベル部分

表1 スタイル情報と言語表現パターン例
Table 1 Style information and expression patterns.

項目	特徴		例
タイトル	S	記事の最上部	
		センタリング	
		記号による飾り	○, □, ★など
		枠囲みなど	
	L	先頭パターン	xx学会，n年度，第n回
		後部パターン	'94
		会議種類パターン	研究会，シンポジウム
		通知パターン1	お知らせ，ご案内，について
		通知パターン2	開催，参加募集，論文募集
開催期日	S	箇条書	
	L	箇条書ラベル	日時，開催日，期日，日程
		日付パターン	n月m日
		接続パターン	～
開催地	S	箇条書	
	L	箇条書ラベル	場所，開催地，会場，ところ
論文締切	S	箇条書	
	L	箇条書ラベル	論文申し込み，締切
		日付パターン	
記事種別	L	会議告知パターン	開催，参加募集，参加者募集
		論文募集パターン	論文募集，論文締切

この表において，Sはスタイル情報，Lは言語表現パターンを表す．

☆ 左端に，タブが存在する行．
☆☆ [*, -, =] などの記号が一定個数以上続く行．

表2 タイトルの言語表現パターン
Table 2 Expression patterns of titles.

パターン名	先頭	任意の文字列	会議種類名	後部	パターン1	パターン2
特徴	(省略可)			(省略可)	単独または重複	
例	xx 学会 xx 支部 n 年度 第 n 回		研究会 例会 セミナー シンポジウム	'94	開催 発表募集 参加募集 論文募集	お知らせ お願い ご案内 の件
凡例	自然言語処理に関するシンポジウム					
	電子情報通信学会ソフトウェアサイエンス研究会					
	AI 学会第 21 回ヒューマンインターフェース研究会発表募集の件					

の直後から，日付パターンを探し，最初に見つかったものを開催期日として抽出する．

(c) 開催地を抽出する場合は，ラベル部分を除去した残りが空でなければ，その部分を開催地として抽出する．残りが空であれば，直後の空行でない行を開催地として抽出する．

(d) 論文締切を抽出する場合は，ラベル部分の直後から，日付パターンを探し，最初に見つかったものを論文締切として抽出する．

(e) タイトル抽出は，ステップ 3 でタイトルが見つからなかった場合のみ行う．ラベル部分を除去した後，残りが空でなければ，その部分をタイトルとして抽出し，空であれば，直後の空行でない行をタイトルとして抽出する．

(5) **未発見情報の再探索**

以上の処理で，必要な情報が見つからなかった場合は，記事の先頭行から探索を行い，未発見情報の抽出を試みる．

(a) 日付パターンを探し，最初に見つかったものを開催期日として抽出する．

(b) 日付パターンと同一行に，締切パターンが存在していれば，その日付を論文締切として抽出する．

(c) 例えば，「第 n 回～研究会」のように，先頭と末尾が明確に限定できるタイトルパターンが見つかれば，それをタイトルとして抽出する．

上記のアルゴリズムで，図 1 に示した記事からサマリー情報を抽出した結果を **図 3** に示す．

記事種別	論文募集
タイトル	第 4 回マルチエージェントと協調計算ワークショップ (MACC'94)
開催日	941011–941013
開催地	名称：ラフォーレ那須
論文締切	940825

図 3 日本語会告記事のサマリー抽出実行例
Fig. 3 Example of summary extraction.

4. ダイジェスト編集

各記事から抽出したサマリー情報から，以下の手順で fj.meetings のダイジェストを作成する．

(1) 全記事に対するサマリー情報から，開催期日・論文締切が現在の日時を過ぎている記事，および，開催期日が抽出できなかった記事のサマリー情報を削除する．

(2) 残ったサマリー情報を，以下のようにまとめる．

(a) 会議告知記事と論文募集記事の二つのグループに分ける．

(b) 前者を開催期日，後者を論文締切でソートする．

(3) 上記のそれぞれを，HTML 形式で出力する．この際，オリジナル記事へのリンク（ニュース記事の ID）を埋め込む．

作成されたダイジェストは，WWW (World-Wide Web) のクライアントプログラム (xmosaic 等) を用いて読むことができる．ハイパーテキストの機能を利用することにより，作成したダイジェストから簡単にオリジナル記事を参照することができる☆．自動作成された会議告知記事のダイジェストの例を **図 4** に示す．

☆ オリジナル記事へのアクセスは，クライアントプログラムの持つニュースサーバへのアクセス機能を利用する．すなわち，ユーザが指定した（身近な）ニュースサーバから記事を得ることになる．

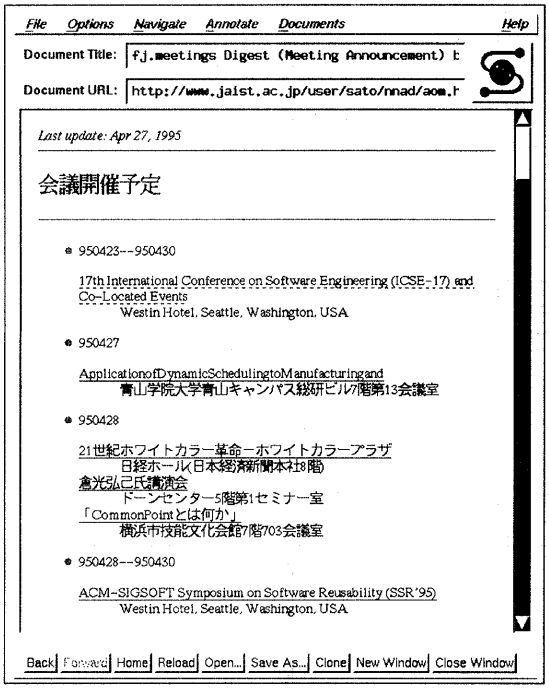


図4 会議開催記事のダイジェスト
Fig. 4 A digest of meeting announcement articles.

5. 実　　験

ダイジェストの自動生成では，サマリー情報をどの位の精度で正しく抽出できるかが，生成するダイジェストの品質を大きく左右する．そこで，その精度を調べる実験を行った．

5.1 日本語会告記事のサマリー抽出実験

日本語会告記事のサマリー抽出実験では，1）サマリー抽出モジュール作成時に調査を行った記事114件（以下，既知データと呼ぶ）と，2）それとは異なる記事97件（以下，未知データと呼ぶ）の2種類のデータに対してサマリー抽出の精度を調べた．なお，後者のデータもfj.meetingsより採集した．

実験の結果を，**表3**に示す．既知データに対しては，各項目の正解率はすべて90%以上を示しており，全体の86%の記事に対しては，すべての情報を正しく抽出することができた．未知データに対しては，各項目の正解率において平均数%程度の低下がみられたが，全体としては，かなり高い正解率を示している．このことより，このサマリー抽出は，ダイジェスト作成に際して，十分実用に耐えると思われる．

未知のデータに対して，比較的大きく正解率が下がった項目は，タイトルと論文締切である．これらの項目の抽出失敗の原因には，以下のものがある．

表3 日本語記事に対するサマリー抽出実験の結果
Table 3 Recall and precision of summary extraction from Japanese articles.

	既知データ		未知データ	
	会議告知 (90記事)	論文募集 (24記事)	会議告知 (73記事)	論文募集 (24記事)
	正解数 (正解率)	正解数 (正解率)	正解数 (正解率)	正解数 (正解率)
記事種別	89 (98.9%)	22 (91.7%)	70 (95.9%)	21 (87.5%)
タイトル	84 (93.3%)	24 (100%)	63 (86.3%)	21 (87.5%)
開催期日	89 (98.9%)	24 (100%)	70 (95.9%)	20 (83.3%)
開催場所	85 (94.4%)	23 (95.8%)	70 (95.9%)	23 (95.8%)
論文締切		22 (91.7%)		19 (79.2%)
総合	77 (85.6%)	21 (87.5%)	57 (78.1%)	13 (54.2%)

(1) タイトルに，想定していなかった表現が含まれていたり，英語表示が混在していた．
(2) タイトルが，空行を挟んだ複数行にわたっていた．
(3) 抽出すべき情報が文章中に埋め込まれていた．
(4) タイトルや論文締切が明示的に表現されていない☆．

このうち，(1)に関しては，タイトルの言語表現パターンを強化することで解決できる．また，(2)に関しては，抽出アルゴリズムを改良することで対処できる．(3)の必要情報が文章中にある場合でも，抽出すべき情報が特徴的な言語表現パターンで書かれている場合は，現在の方法で抽出可能であるが，そのようなパターンが書かれていない場合は，抽出できない．また，(4)の場合は，そもそも抽出すべき情報が明示的に存在しないので，抽出は非常に困難である．

5.2 英語会告記事のサマリー抽出実験

本システムでは，英語会告記事からのサマリー抽出も実現されており，この精度を調べる実験を行った．英語記事に対するサマリー抽出も，日本語記事に対するサマリー抽出と同様に，まず，実際の記事を調査し☆☆，そこから抽出に利用できるスタイル情報と言語表現パターンを整理し，それらを利用した抽出アルゴリズムを作成するという手順を踏んで実現した．基本的なアイディアは，日本語記事に対するものと同じであり，実際に利用するスタイル情報，言語表現パター

☆ 例：「～を目的として集ります」
☆☆ fj.meetingsに流れる英語記事の数が少ないので，comp.ai.*中の会告記事も選び出し，合計104件の記事を採集した．

表4 英語記事に対するサマリー抽出実験の結果
Table 4 Recall and precision of summary extraction from English articles.

	既知データ		未知データ	
	会議告知 (24 記事)	論文募集 (22 記事)	会議告知 (31 記事)	論文募集 (29 記事)
	正解数 (正解率)	正解数 (正解率)	正解数 (正解率)	正解数 (正解率)
記事種別	24 (100%)	22 (100%)	31 (100%)	27 (93.1%)
タイトル	23 (95.8%)	21 (95.5%)	21 (67.7%)	26 (89.7%)
開催期日	23 (95.8%)	21 (95.5%)	27 (87.1%)	28 (96.6%)
開催場所	24 (100%)	21 (95.5%)	26 (87.1%)	23 (79.3%)
論文締切		22 (100%)		23 (79.3%)
総合	23 (95.8%)	19 (86.4%)	17 (54.3%)	15 (51.7%)

ン，抽出アルゴリズムの詳細が異なる．実験結果を**表4**に示す．

この表に示すように，サマリー抽出を実現する際に調査した既知データに対しては，各項目の正解率は非常に高い値を示したが，未知データに対しては，正解率はかなり低下した．特に，低下が著しかった項目は，タイトル，論文締切，開催場所である．これらの原因の多くは，言語表現パターンの不足であり，それは，最初に調査した記事数が46記事と，日本語記事の114記事と比べてかなり少ないことに起因する．この問題は，より多くの記事を調査し，言語表現パターンを強化することによって，日本語記事と同程度のレベルまで改善できると考えられる．

6. 議論および関連研究

6.1 議　　論

(1) 本研究により，電子ニュースのダイジェストの自動生成が可能なことが明らかになった．

電子ニュースにおいては，情報が初めからオンラインテキストとして存在するため，原理的には，ダイジェストの自動生成が可能である．しかし，これまでのダイジェスト☆は，すべて人手によって作成されたものであった．本研究において，fj.meetingsという一つのニュースグループには限定されているが，電子ニュースのダイジェストが自動生成できることを実証した点に大きな意味があると考える．このダイジェストを利用することによって，読者は，記事の取捨選択を容易に行える．

(2) 限定された対象に対しては，表層的な情報だけを利用しても十分な精度でサマリー抽出が行える．

本研究で対象としたfj.meetingsの会告記事においては，スタイル情報と言語表現パターンをうまく組み合わせて利用することにより，サマリー情報をかなり高い精度で抽出することが可能である．このことは，処理対象が十分限定されていれば，このような比較的単純な方法でも，実用的なサマリー抽出が可能であることを示唆している．

(3) ダイジェストの自動生成は，新たな価値を持った情報の自動創出である．

ダイジェストは，単にオリジナル記事のサマリーの集合体であるというだけではなく，それ自体が新たな価値を持った情報である，とみなすことができる．例えば，会告記事のニュースグループのダイジェストは，その時期の会議開催の多少や傾向を示す新しい情報であるとともに，同日に重複開催される会議の一覧を示す情報である．この意味において，ダイジェストは，利用者にとって新しい価値がある情報である．このようなダイジェストを自動生成するということは，新たな価値を持った情報を自動生成するという意味を持つ．

6.2 関連研究

ダイジェスト自動生成システムの中心となる技術は，サマリー情報の自動抽出である．これに関する研究は，主に自然言語処理の分野において，テキストからの情報抽出として研究されてきており，新聞記事からの情報抽出の精度を競うコンテストが毎年開かれている[2]．この他の最近の代表的な研究としては，以下のものがある．

(1) 黒橋らは，専門用語辞典を自動的にハイパーテキスト化するために，リンク，すなわち，用語間の意味関係や用語の説明されている位置などを，文章中の特徴的な言語表現を利用したパターン照合によって自動的に抽出する方法を提案した[3]．

(2) Kameyamaらは，FASTUS[4]と呼ばれる英語パーサを用いて，会議室予約の対話から，会議室予約に関するサマリー情報を抽出することを実現した[5]．

本研究で用いた手法は，黒橋らの手法と同様に，形態素解析を行わずに文字列パターンによって情報を抽出する．単語間に明確な区切りが存在しない日本語テキストでは，テキストを単語に分割する形態素解析が

☆ 例えば，UNIXマガジン（アスキー）には，「NetNews便り」というfjのダイジェストが連載されている．

必要となるが，電子ニュースの記事のように，未編集のテキストや，★や※などのようなノイズが比較的多いテキストに対しては，高い精度の形態素解析は期待できない．このため，形態素解析を行わない本手法は非常に有効である．

本手法のもう一つの特徴は，テキストのスタイル情報を積極的に利用する点である．スタイル情報が多くの有効な情報を担っていることは，これまでも指摘されてきたが，自然言語処理においては，今までさほど積極的に利用されてこなかった．本研究では，ある種のテキストにおいては，スタイル情報がかなりの情報を担っており，それをテキストからの情報抽出に利用できることを明らかにした．このようなスタイル情報の利用は，例えば手紙などのように，明確なスタイルを持つ他テキストからの情報抽出にも利用できると考えられる．

7. おわりに

電子ニュースを「利用しやすい情報メディア」にするためには，そのダイジェストを提供することが不可欠である．本研究では，この電子ニュースのダイジェストを自動生成することを提案し，その一つのプロトタイプとして，会告記事用ニュースグループ fj.meetings のダイジェストを自動生成する方法を示した．ダイジェストの自動生成を実現する中心的な技術は，サマリーの自動抽出技術であり，本研究では，会告記事にみられるスタイル上の特徴，言語表現パターンを利用する方法を提案し，実験によって実用に十分な精度でサマリーを抽出できることを実証した．

本研究で示したダイジェスト自動生成法は，会告記事ニュースグループを対象としたものであり，この方法をそのまま他のニュースグループのダイジェスト生成に適用することはできない．しかし，他のニュースグループに対しても同じようなダイジェスト自動生成を考えることは可能であり，既に，我々は，fj.wanted に対しても，ダイジェスト自動生成システムを実現している[6]．

ダイジェストは，単なる元情報のサマリーの集合体ではなく，新たな価値を持った情報とみなすことができる．この意味において，ダイジェストの自動生成は，情報から新たな情報を生み出す「情報の再生産」を行っているということができ，計算機による情報生産の一形態として注目に値する．

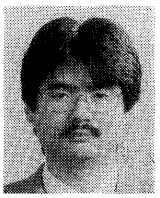
参考文献

1) 佐藤　円：電子ニュースにおけるダイジェスト機構の提案と実現，修士論文，北陸先端科学技術大学院大学情報科学研究科 (1994).
2) Advanced Research Projects Agency: *Proc. of Fifth Message Understanding Conference*, Morgan Kaufmann Publishers (1994).
3) 黒橋禎夫，長尾　眞，佐藤理史，村上雅彦：専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol. 7, No. 2, pp. 336–345 (1992).
4) Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D. and Tyson, M.: FASTUS:A Finite-state Processor for Information Extraction from Real-world Text, *Proc. of 13th International Joint Conference on Artificial Intelligence*, Vol. 2, Morgan Kaufmann, pp. 1172–1178 (1993).
5) Kameyama, M. and Arima, I.: A Minimalist Approach to Information Extraction From Spoken Dialogues, *Proc. of International Symposium on Spoken Dialogue*, Tokyo, Waseda University, pp. 137–140 (1993).
6) 佐藤理史，佐藤　円：ネットニュースのダイジェスト自動生成，言語処理学会第1回年次大会発表論文集，言語処理学会，pp. 297–300 (1995).

(平成7年5月15日受付)
(平成7年7月 7 日採録)

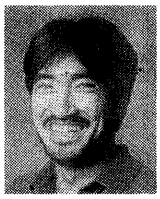
佐藤　円（学生会員）

1986年慶應義塾大学法学部政治学科卒業．同年，(株) 総合ビジョン入社．1990年（株）電通総研勤務．1994年北陸先端科学技術大学院大学情報科学研究科修士課程修了．現在，同博士後期課程在学中．計算機ネットワーク上のマスコミュニケーション，計算機使用者の倫理等に興味を持っている．



佐藤　理史（正会員）

1983年京都大学工学部電気工学第二学科卒業．1988年同大学院博士課程研究指導認定退学．同年，京都大学工学部助手．1992年より北陸先端科学技術大学院大学情報科学研究科助教授．京都大学博士（工学）．自然言語処理，機械学習，超並列人工知能などの研究に従事．人工知能学会，認知科学会，言語処理学会，ソフトウエア科学会，ACL各会員．



篠田 陽一

1983年東京工業大学工学部卒業．1988年同大学工学部情報工学科助手．1991年より北陸先端科学技術大学院大学情報科学研究科助教授．工学博士．分散ネットワークシステムなどの研究に従事．ソフトウェア科学会，ACM各会員．