1 ここまで来たテキスト自動要約

難波 英嗣
広島市立大学情報科学部　nanba@its.hiroshima-cu.ac.jp

奥村　学
東京工業大学精密工学研究所　oku@pi.titech.ac.jp

テキスト自動要約の必要性

インターネットの普及に伴い，電子化されたテキストデータの流通がますます拡大している．このような状況の中，ユーザが必要な情報に効率的にアクセスするのを支援する技術が求められており，さまざまな広がりを見せている．大量のテキストからユーザにとって必要と思われるテキストを検索するテキスト検索技術，大量の情報から重要な情報のみを選択して提供し，要点の迅速な把握を支援するテキスト自動要約技術，ユーザの目的を反映した質問文に対して的確な答えを返す質問応答技術など，さまざまな観点から研究が進んでいる．

本稿では，テキスト自動要約に関する現状を概観する．特に，現在テキスト自動要約関連で研究が活発な技術，注目される話題を紹介するとともに，現在テキスト自動要約技術が実際にどの程度利用可能なのかを実例で紹介し，現状の技術がどの程度のものなのかを考察したい．

最近ますます要約が我々の身近で活躍する場面は増えてきている．近年検索エンジンが広く利用されるようになっているが，システムが提示する検索結果には，（リンク先の）Webページ（テキスト）の内容を短く紹介したものが合わせて提示される場合もある．これは，リンク先のページが，ユーザの欲しいものかどうかを要約を見て判断してもらおうという趣旨でつけられている．また，インターネット上や，携帯電話への配信サービスもある，ニュースの文字放送では，ニュースの原文自体ではなく，その要約といえるようなかたちでニュースが配信されている．さらに，市販のワープロソフトの中には，要約機能を付けたものもかなり見受けられるようになってきた．

このように，テキスト自動要約は現在，人間の知的活動を支援するための基礎技術として重要視され，（本当に「実用的」かどうかは別にして），実用的な自然言語処理（応用）の研究分野の１つとなっていることは間違いないであろう．

これまでのテキスト自動要約

従来どのような手法がテキスト自動要約では用いられてきたのか．この問いへの回答となる解説はすでにいくつか書かれているので，伝統的なテキスト自動要約手法や，その研究動向に関しては，これまで書かれてきた解説等へのポインタを参照しつつ概説することにする．

テキスト自動要約の手法としては，伝統的に，テキスト中の重要な文を抜き出す重要文抽出法が用いられてきた．1990年代に入り，研究が活発化するとともに，その研究の方向性も多様化し，以下のようなトピックスがテキスト自動要約手法に関する研究では注目されるようになっている．

(a) 文から文中の重要個所を抽出することによる要約手

法,
(b) 単一テキストではなく，複数テキストを対象にした要約手法,
(c) 統計的手法あるいは機械学習に基づく要約手法,
(d) ユーザに適応した要約を動的に作成する要約手法,
(e) 言い換え (paraphrase) や書き換え (revision) を用いた要約手法,
(f) 冗長性の少ない要約を目指す要約手法

(a) は，テキストから重要文を抽出する伝統的手法と組み合わせて使用することで，現在の要約システムではほぼ定着しつつある手法といえる．(c) は，近年の自然言語処理研究全般における統計的手法，機械学習に基づく手法の隆盛の一端ともいえるかもしれないが，要約タグ付コーパスを用意することで，自動要約システムを (比較的容易に) 構築できることから，有望視される手法であり，今後も活発に研究が続くものと思われる．(e)，(f) は，単一テキストを対象にした要約において，より自然な要約作成に向けての動きということができ，引き続き研究は活発と思われる．

現時点までのテキスト自動要約に関する研究動向，伝統的な要約手法の詳細に関しては，これまでの解説[6), 7), 11)]を参照していただければ幸いである．また，今はなきbit誌にもテキスト自動要約に関する連載を以前書かせていただいた (テキストを自動的に要約する技術 (その 1-3)，bit,Vol.32, No.2-4, 2000)．テキスト自動要約の入門となることを心がけたつもりなので，この分野で研究を始める人にはまずこちらをお勧めする．

また，自動要約に関する (英語の) 教科書も出版されている[3)]．自動要約に関する話題を分かりやすく記述してあり，この本もこの分野で研究を始める人には必読といえる．なお，この本の翻訳は来年には (共立出版から) 出版されると期待している．

本稿では，(b) を現時点で (最も) ホットなトピックとして後ほど紹介する．残念ながら，この分野での研究はアメリカの方が日本に比べ活発であることは認めざるを得ないところがある．アメリカではDARPA (政府) 主導のプロジェクトでテキスト自動要約研究が進められており，そこから多くの成果が生まれている．プロジェクトの 1 つの方向性には，テキスト集合の情報を簡潔にまとめたテキストの作成，すなわち複数テキストを対象にした要約があり，ここ数年で注目すべき研究がいくつか生まれている．後ほど，それらについて紹介する．

単一テキストを要約する場合に比べ，圧倒的に要約率 (元のテキストの長さと比べた要約の長さ) が小さい複数テキスト要約は，技術的に困難な部分も多く，未解決の課題も多いが，世の中のテキストの量の増大がとどまることを知らず，もはや要約とはいえ，1 つ 1 つのテキストを読み手が読んではいられない状況では，複数テキスト要約技術は不可欠と言わざるを得ないこともこの技術が注目される 1 つの理由といえよう．

最後に，これまでのテキスト自動要約研究の動向においてもう 1 つ特筆すべき点は，自動要約システムの評価方法に関する真摯な議論が継続されている点である．これは，アメリカ，日本両方において行われている．その試みに関しては，本特集の福島らの記事を参照していただきたい．

テキスト自動要約でホットな話題

本章では，テキスト自動要約の分野でも，近年特に活発に研究されている複数テキスト要約について述べる．要約対象が複数の場合，単一テキストの要約の場合と比べ，新たに考慮しなければならない点がいくつかある．本章では，まず，複数テキスト要約のいくつかのポイントについて述べる．次に，これまで開発されてきた技術について簡単に紹介する．最後に，単一テキストから複数テキストへと要約対象が広がっているのと同様，要約対象のテキストのジャンルに関しても，幅の広がりの傾向が現在見られる．その動向に関しても簡単に触れる．

◎複数テキスト要約のポイント

(1) 要約対象テキストの収集

単一テキストの要約と異なり，複数テキスト要約では，要約対象となるテキストをどのように集めてくるかが問題となる．複数テキスト要約の研究が始まった当初は，要約対象として，あらかじめ人間が用意した比較的小規模なテキスト集合が用いられることが多かったが，近年では，情報検索システムの検索結果やテキストを自動分類した結果を直接要約システムに用いるなど，より大規模なテキスト集合を対象とした，実用性の高いシステムが構築されてきている．また，複数のジャンルや複数の言語で書かれたテキスト集合を対象にした要約システムの開発も行われている．

(2) 要約率

複数テキスト要約で要求される要約率は単一テキストに比べ非常に小さい．単一テキスト要約では，一般的には 10 ～数 10% 程度の要約率が設定されていることが多いが，同程度の要約率を複数テキスト要約に適用すると，出力される要約は非常に長くなるので，通常は，単一テキスト要約よりもより小さい要約率が要求される．このため，複数テキスト要約は単一テキスト要約と比べ，より高度な要約技術が必要となる．

(3) 要約対象テキスト間の共通点と相違点の抽出

複数テキスト要約でも，単一テキスト要約と同様に，個々のテキストの重要点を抽出する処理が必要である．しかし，収集してきたテキスト間で内容が重複

する場合，従来の単一テキスト要約の手法を個々のテキストに適用しそれらを並べただけでは，個々の要約の記述が重複し冗長な要約になってしまう可能性がある．このため，冗長な個所（テキスト間の共通個所）をどのように検出し削除するかが問題となる．しかし，冗長な個所を削除しても複数テキストの要約としてはまだ十分であるとはいえない．複数のテキストを要約するとは，それらのテキストを比較し要点をまとめることであるため，テキストの共通点だけでなく，相違点も明らかにする必要があると考えられる．たとえば，ある事件に関する一連の新聞記事から要約を作成する場合，ある記事からは，それよりも日付が前の記事に書かれていない新情報を抜き出し要約に含める必要があるが，この新情報がテキスト間の相違点になっており，どのように抽出するかが課題になる．

(4) 抽出された情報の構成

単一テキスト要約では，抽出された重要個所は，多くの場合，それらが原文中で出現する順番にならべて出力される．しかし，複数テキスト要約の場合，抽出された個々のテキストの重要点，テキスト間の共通点と相違点をどのように構成するかも，考慮すべきポイントの1つになる．要約の構成方法は，対象テキスト集合の性質と関連があると考えられる．たとえば，ある事件に関する一連の報道記事から要約を作成する場合，抽出された情報は基本的には時間順（記事の日付順）に並べればよいと考えられる．しかし，たとえば「2000年問題に対する各社の対応状況」に関する要約を作成する場合，各社の対応状況はそれぞれ独立したイベントであるため，それらを時間順に並べるのが最適な構成方法であるとは限らない．

(5) 要約の読みやすさ

単一テキスト要約（特に重要個所抽出型の要約）では，抽出された部分テキスト（パッセージ）を単純に並べると，それらが原文中で離れた個所から抽出されている場合には，その個所で文間のつながりが悪くなる可能性がある．このような場合には，適宜接続詞を補ったり，不要な接続詞，照応詞等を除去したり，また，抽出された個所中に宙ぶらりんな（先行詞がない）照応表現が存在する場合はその表現を先行詞に置き換える，といった処理が必要になる．同様の処理は複数テキスト要約でも必要であるが，複数テキスト要約の場合，さらに考慮すべき点がある．たとえば，テキスト間で文体が異なる場合には，文体を統一した上で要約を作成する必要がある．また，前で述べた「抽出された情報をどのような順序で出力するか」は，文間のつながりの良さという点で，要約の読みやすさとも関連があると考えられる．

(6) 要約の提示方法

複数のテキストに書かれている多くの情報を限られたスペースでユーザに効率的に伝達するために，さまざまな方法が考えられる．たとえば，文書としてまとめるのも1つの方法ではあるが，このほかにもいくつかの項目をリストや表などにまとめるといった方法も考えられる．

◎これまでの取り組みと開発技術

ここでは，先に述べた複数テキスト要約のポイントの中でも，特に (3) と (4) を取り上げる．(3) はこれまでにこの分野でしばしば議論されてきており，またさまざまな方法が提案されている．また (4) は (3) と比べるとまだ十分議論されているとはいえないが，この分野で関心を集めているトピックの1つである．

(3) 要約対象テキスト間の共通点と相違点の抽出

テキスト間で共通する個所を同定する方法がこれまでにいくつか提案されている．それらは，形態素レベルの比較と構文レベルの比較の2つに大別できる．異なるテキスト中に存在する以下の2文を同定する処理を例に，これらの2つの代表的な方法を説明する．

[文 1]「フーバー社は軽量の携帯電話を発売する」
[文 2]「軽量の携帯電話が来月フーバー社から発売される」

形態素レベルの比較[☆1]では，まず比較する2文を形態素解析し，そこから自立語を抽出する．上の例では，文1からは（フーバー社，軽量，携帯電話，発売）が，文2からは（軽量，携帯電話，来月，フーバー社，発売）が，それぞれ抽出される．これらの語の中で，文1では4語中4語が，文2では5語中4語が互いに一致している．このように一致の度合が高ければ，2文は同じ内容について述べていると考え，この2文をテキスト間の共通個所として同定する．

これに対し，近年の構文解析技術の向上に伴って，構文レベルで比較しようという試みもある[1), 4), 12)]．この方法では，あらかじめ比較対象の2文を構文解析しておき，解析結果として得られた構文木をいくつかの規則を用いて変形し，2文が同一内容であるかどうかを同定する．さらにこれらを統合して1つの文を生成する．文1と文2の例では，文2が文1の受動態になっているため，能動態を受動態に変形する規則を作成しておけば，文1と文2の同定が可能になる．また，文2には「来月」という文1にはない語が含まれているが，構文解析の結果「来月」が「発売される」に係っていることが分かれば，文2に能動態変形規則を適用し「来月」が「発売する」に係るよう変形した上で文1との統合処理を行えば，以下のような文3を生成することが可能である．

☆1 より詳しい解説は文献6）を参照していただきたい．

［文 3］「フーバー社は軽量の携帯電話を来月発売する」

(4) 抽出された情報の構成

人間が複数テキスト要約を作成する際，さまざまなトピックの並べ方が存在するが，そこには何らかの制約が働いていると考えられる．Barzilayら[2]は，複数の記事から抽出されたいくつかの重要文のセットを 10 人の被験者に与え，それらを並べ換えることで要約を作成してもらっている．そして，その結果を比較することで，次のような知見を得ている．すべての文の順序が被験者間で完全に一致することはあまりない．しかし，順序が入れ換わっても，常に隣り合って出現する文のペアがいくつかある．これらのペアは関連したトピックの文で構成されている．したがって，複数テキスト要約において文間の結束性を考慮することは重要である．

このような知見に基づき，Barzilayら[2]は，要約文の順序を決定する方法を考案している．基本的にはトピックを時間順に並べるが，関連したトピックの文は必ず隣接して出力する．この方法により，作成される要約文書はある程度一貫性が保たれる．Barzilayらは，この手法を先に述べた「記事が書かれた時間順に並べる方法」と比較し，前者の手法の方が優れていることを示している．

なお，この手法に基づいた要約システムを後ほど紹介する．

◎要約対象の幅の広がり

これまでの自動要約研究の多くは，その要約対象のテキストのジャンルとして，新聞記事，論文を扱ってきた．これに対し，近年これ以外のジャンルのテキストを要約対象とする研究が見られるようになってきた．たとえば，Webページを対象とした研究や，電子メールを対象とした研究がある．

さらに，テキストではなく，音声（あるいは，その書き起こしである話し言葉のデータ）を対象とする要約研究がいくつか見られるようになってきた．これには，講演音声のようなmonologueと，2 人以上による対話（dialogue）の両方が含まれる．話し言葉を対象とした要約では，

- テキストとしての情報以外に他の音響的情報が利用できる，
- 音声認識結果を入力とすることから，入力にノイズが含まれる，
- 話し言葉の特性としての冗長性が入力には含まれる，

など，テキストを対象とした場合とは異なり，新たに考慮しなければいけない点が存在する．

本節のトピックの詳細は文献 7）に譲るが，このように多様な種類の入力を自動要約システムで扱おうとする試みは，後でも述べるように，自動要約システムがアプリケーション指向で今後開発されていくであろうことを考えると，今後も続くと思われ，注目されるトピックということができる．

使われているテキスト自動要約

本章では，テキスト自動要約技術を利用した，商用ソフトウェア，利用に供されているサービスを可能な限り紹介し，現状の技術がどの程度実用可能なのかを考察する．なお，ここで挙げるものは，筆者が知る範囲で特徴的なものに限定しており，網羅的なものではないことをお断わりしておく．

市販のソフトの中で，要約だけを専門に行うソフトは，現在のところ恐らく販売されていないと思われる．しかし，テキストを扱う市販ソフトに要約機能が付いているケースはたくさんある．専門ソフトにはなっていないものの，他のソフトの基本機能として要約機能が付くのは，もはや当たり前のことになったといえる．その代表的なものがワードプロセッサ（ワープロ）ソフトであろう．要約機能が付いている代表的な日本語ワープロソフトとしては，マイクロソフトワード，富士通OASYS，ジャストシステム一太郎，ロータスワードプロの 4 つがある．これらのソフトの要約機能の詳細はここでは省略するが，先に紹介したbit誌の連載に詳細があるので，興味のある方はそちらを参照していただきたい．

ワープロソフトのほかに，最近メジャーになってきたテキストデータベースソフトや，翻訳ソフトにも要約機能付きのものがある．また，最近ではテキストデータベースに近い関係にあるテキストマイニングという領域のソフトに，さまざまなツールの 1 つとして要約機能が提供されている．たとえば，IBMのIntelligent Miner for TextのSummarization Toolは単語の重要度と文の重要度の 2 段階の処理で重要文を抽出している．現在要約機能の付いていないソフトにも，バージョンアップなどのおりに新たに追加されるなど，要約機能付きのソフトは今後ますます増えていくものと思われる．

また，Alta Vista Discovery検索エンジンのインタフェース中には，InXight社の要約器が搭載されているなど，検索エンジンの出力表示に要約を用いるというのも現在ではもう珍しくないかもしれない．

本章の残りでは，Web上で利用可能な 2 つの複数テキスト要約システムを，先に述べた作成のポイントに沿って紹介する．どちらもWeb上の英文ニュース記事

を対象にしている．GoogleでもGoogle News（http://news.google.com/）というサービスを始めており，複数のニュース記事をまとめて1つの要約として作成する，これらのシステムは，Web上で実際に利用可能であることも含め，大変興味深いといえる．

◎Newsblaster

Newsblaster（http://www.cs.columbia.edu/nlp/newsblaster/）は，Columbia大学のMcKeownらのグループが開発した要約システムである[5]．このシステムは，CNN, Reuters, Fox News, NY Post, USA Today等の17のニュースサイトから新聞記事を収集し，要約を自動作成する．本節では，まずNewsblasterの動作例を示し，次にシステムの構成について説明する．

• Newsblasterの動作例

図**-1**は，Newsblasterのトップページ（2002年10月13日現在）である．要約は「U.S.」，「国際（World）」，「金融（Finance）」，「娯楽（Entertainment）」，「科学技術（Science/Technology）」，「スポーツ（Sports）」の6つのカテゴリーにあらかじめ分類されている．これらのカテゴリーは図中でグレーの帯で示されている（実際のシステムではオレンジ色で表示される）．各カテゴリー内で，関連するイベントクラスタ（記事集合）はグルーピングされており，グループごとにキーワードが示してある．たとえば，図のFinanceというカテゴリーの1つ目は，「Dow, Commerce Department, Dow Jones, Wall Street, Iraq」というキーワードに関連するイベントクラスタが2つあり，それぞれ4つの記事を含んでいることを示す．

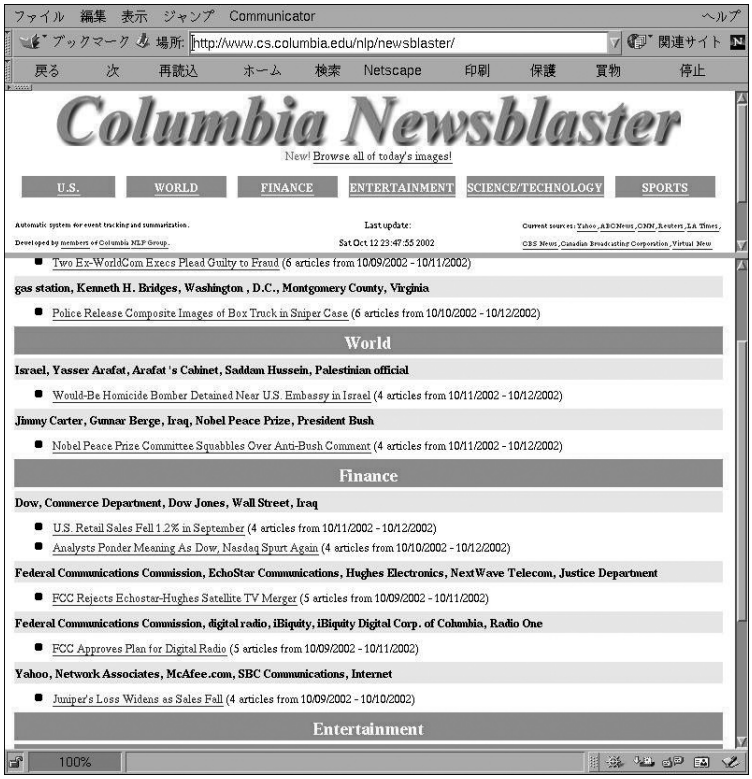

図 **-1　Newsblaster** 動作例 **1**
http://www.cs.columbia.edu/nlp/newsblaster/archive/353.html

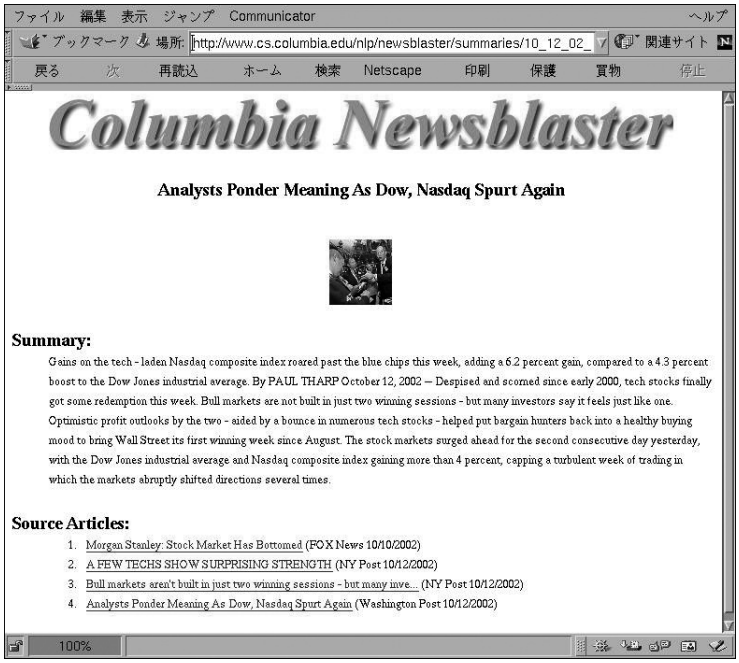

図 **-2　Newsblaster** 動作例 **2**
http://www.cs.columbia.edu/nlp/newsblaster/archive/summaries/353/10_12_02_7.html

このうち「Analysts Ponder Meaning as Dow, Nasdaq Spurt Again」のリンクをクリックすると，4つの記事から生成された要約が表示される（図**-2**）．この要約のタイトル「Analysts Ponder Meaning as Dow, Nasdaq Spurt Again」は，4記事の中で最も代表的な記事（Washington Post）のタイトルである．また，タイトル・要約と一緒に，関連する画像のサムネイルが表示されている．要約の下には4記事へのリンクが張ってあり，原文を参照できるようになっている．

次に，システムの構成について説明する．

システムの構成

Newsblasterは以下に示すパイプラインアーキテクチャに従っている．

1．17のニュースサイトから新聞記事を収集．画像を抽出．
2．収集した記事をイベントごとにクラスタリング．
3．イベントクラスタを6つのカテゴリーに分類．
4．関連する複数のイベントクラスタをグループ化．
5．イベントクラスタごとに要約を作成．

以下，1．～5．について順に説明する．なお，先に述べた複数テキスト要約のポイントと関連する項目には (1) ～ (6) を合わせて記す．

1．17 のニュースサイトから新聞記事を収集．関連画像を抽出：
Newsblaster は 17 のサイトを巡回し，トップページから最大 4 階層までたどって，Web 文書を収集する (1)．次に，Web 文書から新聞記事を抽出する．記事の抽出には次のようなヒューリスティクスを用いている．Web 文書中の最も大きなセルの中で 512 文字を超えるものがあれば，それを新聞記事と考え抽出する．また，関連画像は「記事と同一セル中にあり，かつ画像の URL 中に ad や advertisement を含まない」といった情報に基づいて抽出する．

2．収集した記事をイベントごとにクラスタリング：
1．の過程で収集された記事について，記事に含まれる固有名や用語等を用いて，イベントごとにクラスタリングを行う (1)．

3．イベントクラスタを 6 つのカテゴリーに分類：
イベントクラスタは，6 つのカテゴリー「U.S.」，「国際 (World)」，「金融 (Finance)」，「娯楽 (Entertainment)」，「科学技術 (Science/Technology)」，「スポーツ (Sports)」のいずれかに分類される．あらかじめ，カテゴリーごとに TF・IDF という手法を用いて，カテゴリーの特徴を表す用語ベクトルを作成しておく．次に，イベントクラスタ中の各記事とカテゴリーの用語ベクトルを比較し，各記事に最も近いカテゴリーを割り振る．イベントクラスタ中で，最も多くの記事が割り振られたカテゴリーにイベントクラスタを分類する．

4．関連する複数のイベントクラスタをグループ化：
同じカテゴリーに分類された複数のイベントクラスタの中で，関連するものは，2．で用いられたクラスタリング手法を用いてグループ化される．図-1 に示すように，各グループにはグループの特徴を示すキーワードが示されている．これらは，まずクラスタ中のすべての記事から固有名詞と用語を抽出し，出現頻度と IDF により重み付け，次にそれらの上位 5 つを関連イベントの最も代表的な用語として選択している．

5．イベントクラスタごとに要約を作成：
イベントクラスタ中の記事は，Columbia Summarizer に送られ，要約が生成される．Columbia Summarizer とは，以下に示す 4 種類のクラスタのタイプを考慮し，タイプごとに異なる要約方法を用いる混合式要約システムである[☆2]．

(I)「単一イベント文書」は，ほぼ同時期にある場所で起きた 1 つのイベントに焦点を合わせている．

(II)「人物に焦点をあてた (伝記) 文書」は，ある人物に関するイベントや，その人物に関する背景情報を取り扱う．

(III)「複数イベント文書」は，異なる場所・時間，また普通は異なる主唱者によるいくつかのイベントについて言及する．

(IV)「その他クラスタ」は，大まかに関連する文書集合を含んでおり，先に述べた 3 つのカテゴリーに含まれない．

Columbia Summarizer 内部では，まずルータがクラスごとの文書タイプを決定し，次に適切な要約サブコンポーネントを起動する．

(I) に関しては，MultiGen というシステムを用いて要約の作成を行う．MultiGen は，「複数テキスト要約のポイント」(3) で説明したように，異なる記事中の 2 文を，まず語 (形態素) レベルで比較し，類似する文を同定しておく．次にこれらを構文レベルで比較・統合して 1 つの文として出力する．統合された文を，「複数テキスト要約のポイント」(3) で説明した方法で順序付け，結果を出力する．

(II)（III）(IV) に関しては，DEMS というシステムを用いる．DEMS は，記事の第 1 段落に多く含まれる単語，入力記事中の概念 (WordNet 中の特定の意味的なグループ) の頻度，記事の日付 (新しい記事の情報が重要)，代名詞の有無 (宙ぶらりんな照応が要約に含まれるのを避けるため，代名詞を含む文は負の重みが与えられる) や文の長さ (短すぎる文は曖昧で，長すぎる文は無関係な情報を含む) といったいくつかの素性を組み合わせて，重要な情報を含んだ文を選択する．また (II) に関しては，このほかに文がターゲットの人物名を含んでいるか，という情報も合わせて用いている．

(II)（III）(IV) のような記事集合から抽出された文をどのような順序で出力するかについては，まだ十分に議論されていないようである．しかし，出力される要約の冗長性を抑えるため，要約中の繰り返し文のチェックは少なくとも行っている (3)．

◎NewsInEssence

NewsInEssence（http://www.newsinessence.com）は，Michigan 大学の Radev らが開発した要約システムである[9), 10)]．このシステムは BBC, CNN, MSNBC, USA Today, Yahoo! の 6 つのニュースサイトから新聞記事を収集し，要約を自動作成する．本節では，まず NewsInEssence の動作例を示し，次にシステムの構成について説明する．

[☆2] ただし，実際には（III）と（IV）は，同じ要約器を用いて要約を生成しているので，厳密には 3 種類のタイプを考慮していることになる．

NewsInEssenceの動作例

図-3 は，NewsInEssenceのトップページ（2002 年 10 月 13 日現在）である．トップページでは，NewsInEssenceがニュースサイトを巡回して収集し，作成した要約のうちの 1 つを表題（この場合，"CNN.com - Kuwait says al Qaeda linked to attack on Marines - Oct.12, 2002"）とともに示している．図-3 中のFULL SUMMARYという個所をクリックすると，この要約の原文およびさらに長い要約が示される（図-4）．図-4 より，図-3 の要約は 7 つのニュースサイト中の 8 記事から作成されていることが分かる．また，図-4 中の［5%］［10%］［20%］という個所をクリックすると，さらに長い要約が提示される．

NewsInEssenceは，あらかじめニュースサイトから記事を収集して要約をユーザに提示する機能を備えている点はNewsblasterと同様であるが，このほかに，ユーザが関心を持っているニュースについて，インタラクティブに要約を作成する機能もある．トップページには，ユーザが関心を持っている記事（以後，種記事と呼ぶ）のURLや複数のキーワードを入力するためのフォームがあり，ユーザがフォームに記入すると，NewsInEssenceは直ちにニュースサイトを巡回して関連記事を収集，要約を作成し，数分程度で結果を電子メールで通知してくれる．このようにして作成された要約はサーバ側で保存しており，他のユーザが閲覧することも可能である．

次に，システムの構成について説明する．

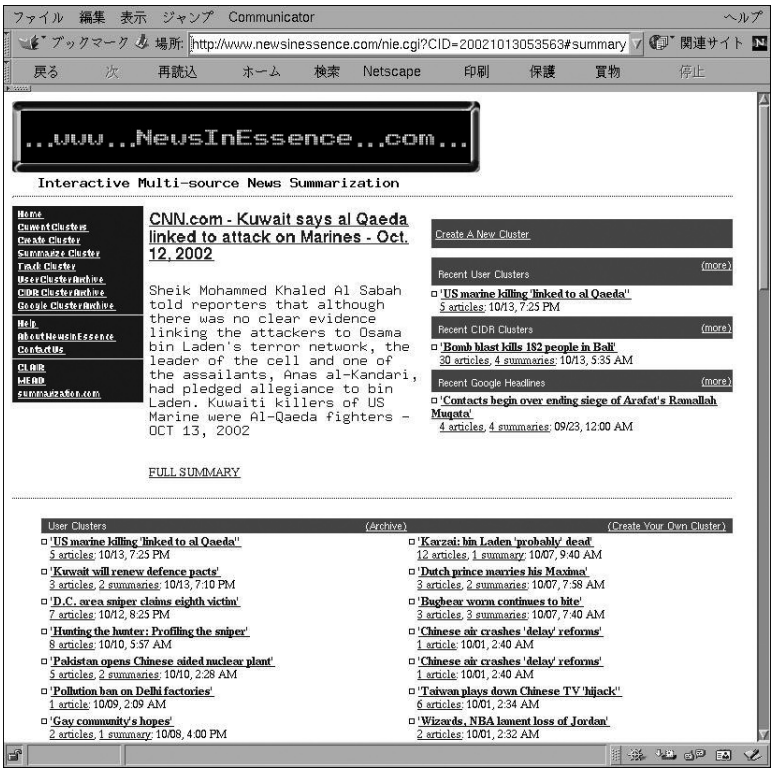

図-3　NewsInEssence 動作例 1
http://www.newsinessence.com/nie.cgi?CID=20021013053563

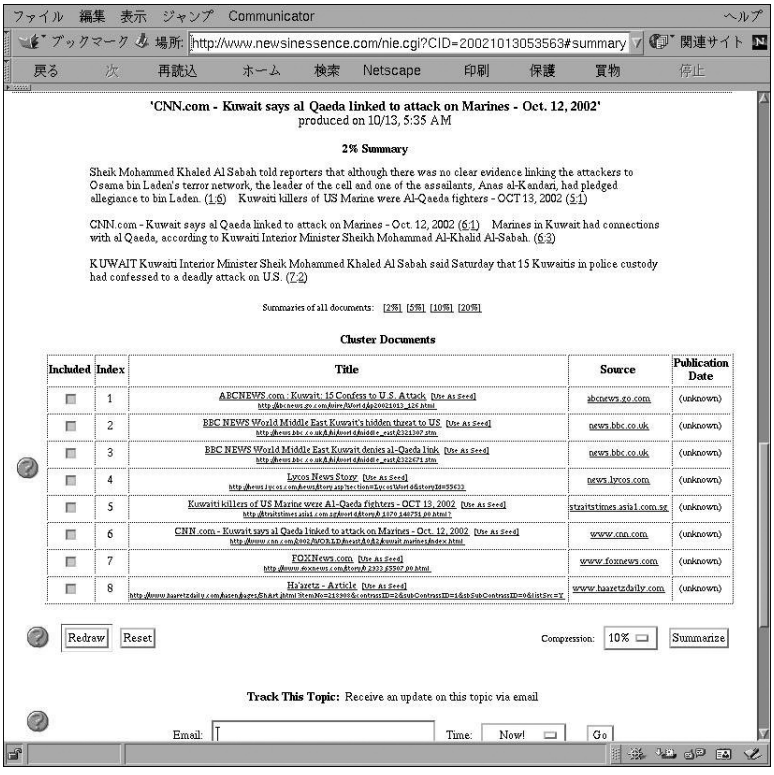

図-4　NewsInEssence 動作例 2
http://www.newsinessence.com/nie.cgi?CID=20021013053563#summary

システムの構成

ここでは，NewsInEssenceが提供するいくつかの機能のうち，特にユーザの種記事を元に要約を作成する場合を取り上げ，その手順について述べる．

NewsInEssenceは，NewsTrollという記事の収集を行うエージェントと，centroid-based summarizationという技術に基づいた要約システムから構成されている．以下では，これらについて順に説明する．

1．NewsTrollによる関連記事の収集：

ユーザが種記事を入力すると，NewsTrollはリアルタイムでインターネットから関連記事を収集する（1）．NewsTrollは 2 段階で記事の収集を行う．第 1 段階では，種記事を含むページからリンクをたどって関連記事を探す（以後，収集された記事集合をクラスタと呼ぶ）．この過程で，クラスタのキーワードを決定する．キーワードの決定には，TF*IDFという手法を用い，クラスタ中にはしばしば出現するが一般的にはあまり出現しない語を選ぶ．

キーワードを選定すると，収集の第 2 段階に入る．この段階では，NewsTrollは第 1 段階で決定したキー

ワードを用い，6つのニュースサイトの検索エンジンを用いて関連記事をクラスタに追加しようと試みる．

追加候補となる記事は，IDFで重み付けられたn次元ベクトルで表現され，種記事との類似度は余弦距離で計算される．もし，類似度がある与えられた閾値以上であれば，そのページは関連記事を含んでいると考え，クラスタに追加される．

2．centroid-based summarizationによる要約作成：

種記事と関連する記事集合のクラスタが形成されると，次にcentroid-based summarizationという技術を用いて要約が作成される[8)]．

まずクラスタに含まれる記事の各文の重要度を計算する．各文の重要度は，「1．で決定したキーワードをその文がどの程度含んでいるか」「各文の記事中での位置」等の情報に基づき計算される．次に要約率に応じてスコアの高い順に文を抽出しそれらを要約に含めるが，このままでは類似した内容の複数の文が抽出される可能性があり，結果として出力される要約も冗長になる．そこで，ある文が，それより高いスコアを持つ文と，語のオーバーラップがある場合には，冗長性ペナルティを与え，その文のスコアを下げることで対処している（3）．抽出された文は，文を含む記事の書かれた日付順に並べ（4），それらが最終的に要約として出力される．

どこへ行くテキスト自動要約

本稿では，テキスト自動要約研究の現状およびその現在の技術水準について概観した．テキスト自動要約技術の現状を踏まえ，今後この分野の研究はどのような方向へ進むのかであるが，自動要約研究が応用に近い研究分野ということもあり，今後ますます要約の用途，利用目的を特化したかたちでの要約手法の研究が活発化すると思われる．

テキスト自動要約技術の応用として，いくつかの新しい方向性が明確になってきたこともここ数年の特徴といえる．これまでも，サーチエンジンにおける検索結果の表示や，ユーザのナビゲーションにおいて要約を利用する研究や，字幕作成，文字放送用に要約手法を利用することは試みられていた（たとえば，本特集の江原らの記事を参照）．これに加えて，ここ数年で，携帯端末における情報提示のための要約の利用（たとえば，本特集の中川らの記事を参照）や，（高齢者，視聴覚障害者といった）情報弱者のための情報保証への要約の利用（たとえば，自動要約筆記やユーザの視覚特性に合わせたトランスコーディング）といった，新しい有望な応用分野が要約には付け加わったといえる．

一方，今後この分野の研究はどのような方向へ進むべきなのかであるが，上で述べた応用指向の方向性と相反するように見えるかもしれないが，これまでより精緻な言語処理を利用した要約技術に関する方向性が重要であると考えている．先に述べたように，自動要約研究の方向性のうち，いくつかは，人間の作成した要約と同様の要約を目指すものがあり，このためには，現在研究が途上の段階にある，文脈解析技術，言い換え技術，書き換え技術，テキスト生成技術等が不可欠である．これらの要素技術および，それらを用いた要約技術は，もちろん一般の任意のテキストに適用可能なものを開発するのはまだまだ困難であろうが，要約の応用を限定し，要約対象を特化することにより，ある程度実用的なものの開発の見通しが立つように思われる．今後の自動要約研究では，実際の応用の可能性を検討しつつも，その中で，決して表層的な手法に陥らず，ある程度深い解析技術等を利用した要約手法の研究を行っていきたいと，1研究者として，自戒の気持ちも含め，最後に記しておきたい．

参考文献

1）Barzilay, R., McKeown, K.R. and Elhadad, N.: Information Fusion in the Context of Multi-Document Summarization, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.550-557 (1999).

2）Barzilay R., Elhadad, N. and McKeown, K.R.: Sentence Ordering in Multidocument Summarization, Proceedings of HLT'01, pp.149-155 (2001).

3）Mani, I.: Automatic Summarization, John Benjamins Publishing Company (2001).

4）McKeown, K.R., Klavans, J.L., Hatzivassiloglou, V., Barzilay, R. and Eskin, E.: Towards Multidocument Summarization by Reformulation: Progress and Prospects, Proceedings of the 16th National Conference on American Association for Artificial Intelligence, pp.453-460 (1999).

5）McKeown, K.R., Barzilay, R., Evans, D., Hatzivassilogtou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster, Proceedings of HLT'02 (2002).

6）奥村　学，難波英嗣: テキスト自動要約に関する研究動向，自然言語処理，Vol.6, No.6, pp.1-26 (1999).

7）奥村　学，難波英嗣: テキスト自動要約に関する最近の話題，自然言語処理，「自動要約」特集号，Vol.9, No.4, pp.97-116 (2002).

8）Radev, D.R., Jing, H. and Budzikowska, M.: Centroid-based Summarization of Multiple Documents: Sentence Extraction,Utility-based Evaluation and User Studies, Proceedings of ANLP/NAACL 2000 Workshop: Automatic Summarization, pp.21-30 (2000).

9）Radev, D.R., Blair-Goldensohn, S., Zhang, Z. and Raghavan, R.S.: Interactive, Domain-independent Identification and Summarization of Topically Related News Articles, Proceedings of the Fifth European Conference on Research and Advanced Technology for Digital Libraries (2001).

10）Radev, D.R., Blair-Goldensohn, S., Zhang, Z. and Raghavan, R.S.: NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization, Proceedings of HLT'01 (2001). http://perun.si.umich.edu/~radev/papers/hltd01.ps

11）佐藤理史，奥村　学: 電脳文章要約術－計算機はいかにしてテキストを要約するか－，情報処理，Vol.40, No.2, pp.157-161 (Feb. 1999).

12）上田良寛，小山剛弘: 共通意味断片の抽出による複数文書要約，言語処理学会 第6回年次大会，pp.360-363 (2000).

（平成14年10月25日受付）