類似意味内容の統合による伝達型電子化文書要約方式の提案

4R－11

稲垣 博人　早川 和宏　田中 一男

NTTヒューマンインタフェース研究所

1 はじめに

電子メイル，web等により，大量のデジタル情報が流通される時代となった．これらの大量なデジタル情報を効率よく処理し，吸収し，自分のものとするためには，情報を効果的に提示したり，情報を要約表示する必要がある．

情報の要約を抽出するためには，情報の中に記述されている重要な語や内容を表す重要な事柄を抽出する必要がある．重要語により要約を生成する手法としては，キーワード抽出と同様な手法[1]により，重要語を抽出し，その重要語が多く含まれている文を要約とする手法がある．一方，文書に記述されている事柄を抽出することにより，重要なメッセージを抽出する手法[2]も考えられる．しかし，人それぞれ興味対象にばらつきがあるため，一概に重要な情報を抽出することは難しい．

そこで我々は，同一の事柄について記述している電子化文書に対して，複数の人間が重要であると考えている情報の和集合を抽出し，その和集合を電子化文書の要約に利用する要約方式を提案する．例えば，新聞記事のように，あるニュースソースの情報に基づき，各社が独自の調査，情報を付加・削除をおこなうことにより生成される伝達型文書では，同一の事柄について記述しているものの，記述する内容・表現に“ばらつき”がある．このような記事の特徴を利用して，“ばらつき”の少ない表現を中心に集めて要約を生成する方式である．

2 類似意味内容の統合による伝達型文書要約

本稿で提案する類似意味内容の統合による伝達型文書の要約方式は，同じ事柄について表現された複数の伝達型文書の中で，同じ事柄について述べられている情報をピックアップし，その情報の和集合を要約として構成する方式である．

まず，類似な意味内容が記述されている文書を文書群からピックアップする．次に，それらの文書に記述されている内容を比較し，類似する意味内容の情報を抽出する(図1参照)．

これらの処理の自動化のために，以下の項目を検討する．

- 文書情報の構造化処理

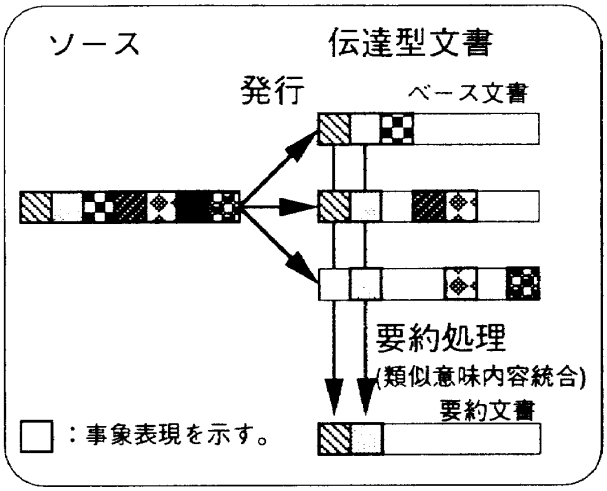

図1: 伝達型文書の要約

- 類似文書および類似意味内容の抽出処理
- 類似意味内容の比較による要約の構成処理

ここでは，伝達型文書として，新聞記事を対象とした．

2.1 文書情報の構造化処理

個々の文書に記述されている意味内容を比較するためには，文書がどのような意味内容を記述しているかを的確に構造化しなければならない．

ここでは，“事象”[3]という概念に基づいてメッセージを構造化する．“事象”とは，「世界に発生する事柄の変化」である．その事柄の変化を言葉で表現する要素が事象要素である．必須の事象要素としては，“動作主体”，“動作対象”，“動作内容”，“場所（場，起点，終点）”，“時間（場，開始時刻，終了時刻）”などの5要素がある．さらに，これらの5要素を修飾する事象修飾要素がある．

まず，入力された文を形態素解析，係り受け解析した後，事象における“変化”の部分である“動作内容”表現をピックアップするとともに，その“動作内容”に付随する各事象要素を文中から抽出する．事象の修飾要素は，事象要素に対して修飾関係にある表現や，「は」「と」「を」などの格助詞などの手掛かり表現をもとに抽出される．各事象要素には，該当するパッセージ（単語や句など，語の組合せをいう．）が記述される．

例えば，「NTTがA社の交換機を購入した」という文では，表1のような事象構造が抽出される．

A Proposal of the message abstraction method utilizing similar passages of each news from one news source.

Hirohito INAGAKI, Kazuhiro HAYAKAWA, and Kazuo TANAKA.

NTT Human Interface Laboratories

表 1: 事象構造抽出例

事象名	パッセージ
動作主	“NTTが”
動作対象	“交換機を”
動作対象の修飾	“A社の”
動作内容	“購入した”

2.2 類似文書および類似意味内容の抽出処理

類似文書および類似意味内容の抽出処理では，まず，新聞記事群に対して類似する記事を抽出する．次に，類似する記事の意味内容を解析し，類似意味内容を抽出する．

新聞記事群から類似する記事を抽出するために，荒い記事の絞り込みを行なう．絞り込みはここでは，一般の新聞記事に付与されている統制語，またはフリーキーワードを利用して新聞記事を選別する．例えば，“NTTに関係する情報の中で・・・”というような場合，記事データベースに対して，“NTT”をキーとしてDB検索を実施し，記事群を荒く選別する．

次に荒く選別した記事から，同じニュースソースから発信されたと考えられる伝達型の記事群を抽出する．

伝達型記事では，表2のような規則が成り立つ．

表 2: 伝達型記事のルール

事象発生日時 < ソース発信日時 < 伝達型記事発行日時

一般的に，新聞記事は，重要なニュースや，速報性が求められるニュースの場合，ニュースソース発信日時から，1日内外で，新聞記事として発行される．一方，即時性の求められない記事や，特集記事などの場合，ニュースソースが発信されてから即発行されるわけではなく，誌面の都合等により，発行日時が決定される．ニュースソースの発信から，日が経って発行される場合，複数の情報源からの情報や，記者の主観など，ニュースソースの情報からかなり変遷し，同一のニュースソースからの情報であると判断できない場合もある．

ここでは，ある新聞記事が発行され，その発行から＋1日以内で発行された新聞記事を類似記事の検査対象とする．

荒く絞り込まれた記事群から，ニュースソースが同一の新聞記事を抽出するために，まず，荒く絞り込まれた記事群に対して事象解析を行なう．次に，記事群の中で，まず1記事を選択し，その記事の発行日から＋1日以内の記事の中で，各記事の先頭の数事象が類似しているかを比較する．類似しているかは，事象要素のパッセージを構成する単語の類似度で判断する．

ニュースソースが同一であるとする記事群を抽出した後，各記事の事象解析をもとに，類似する意味内容（類似する事象）を抽出する．類似する意味内容としては，各事象要素が少なくとも3要素以上類似した事象を同一事象と判定する．

2.3 類似意味内容の比較による要約の構成処理

要約の構成処理では，ニュースソースが同一と推定される記事群の中から要約のベースとなる記事（ベース記事）を中心に要約文を構成する．

要約のベース記事は，抽出された記事群の中で，発行日が最も古く，かつ最もコンテンツ量が少ないものとする．これは，発行日が古いものほど，ニュースの初期の重要な情報を伝達しており，コンテンツ量が少ないものほど，重要な情報を用いて表現していると考えられるからである．

上記に適合するベース記事を決定し，要約を生成する．その場合，抽出された記事群の中で，同一の事象が最低でも2つ以上ある事象を採用する．さらに，類似した事象の中で，類似するパッセージだけを要約パッセージ候補として抽出する．

最終的に，要約パッセージ候補をベース記事の語順に基づき再構成し，構成した要約が非文とはならないように，修飾節のピックアップや文末表現の変換を行ない要約文として出力する．

3 まとめ

本稿では，あるソースの情報をもとに伝聞される伝達型文書を要約する手法を提案した．特に，同一の事柄について述べる伝達型文書の代表である，新聞記事を対象とし，同一の事柄に関して記述した記事の事象表現の和集合を抽出・再構成することにより要約を生成する手法である．

今後は，本提案方式を計算機上にインプリメントするとともに，新聞記事の要約精度を評価する．

参 考 文 献

1) H. P. Luhn. The automatic creation of literature abstract. *IBM Journal, Vol2*, 1958.

2) Hirohito Inagaki and Tohru Nakagawa. An abstraction method using a semantic engine based on language information structure. *Coling-92*, 1992.

3) 稲垣博人. 事象解析による要約情報の抽出. 情報処理学会自然言語研究会,NL84-3, 1991.