データベースシステム　127－19
情　報　学　基　礎　　67－19
（２００２．５．２２）

時系列データを入力とした文生成技術に基づく複数言語要約システム

関洋平
seki@it.aoyama.ac.jp
青山学院大学理工学部

本研究は，言語に応じた要約処理についての手法について提案する，複数言語要約とは，本研究では単一言語を入力として複数言語を出力するものを対象とする．要約は文抽出技術に基づいて構成されるものが多いが，複数言語を出力するためには，文を単位として抽出を行う場合，出力に機械翻訳システムを適用することになる．本研究では，時系列データを入力として文章を生成する技術に基づいて，時系列に基づいてテキストデータを整理することにより，複数言語要約を実現する．入力は日本語の相場市場に関する新聞記事を採用し，日英二ヶ国語の要約を生成する．

Multilingual Summarization System Based on NLG Technology from Time-Series Data.

Yohei SEKI
seki@it.aoyama.ac.jp
Aoyama Gakuin University

I propose a method to implement multilingual summarization in this paper. Multilingual Summarization means here summary produced with several languages from one language document. Although summarization techniques today are mainly based on sentence extraction and revision method, summary construction units in my research are extracted based on chronological data sets and my system produces multilingual summary from those units with language dependent discourse structure. This research concerned with producing Japanese and English summary from Japanese newspaper articles about the market price.

1 はじめに

本研究は，単一言語の複数文書を入力として，複数言語の要約生成を試みる技術の一つを提案する．複数言語の要約を生成するためには，各言語ごとに文章を生成するために必要な情報について詳細に定式化する必要がある．文章を生成する技術には，文章全体の談話構造を定式化して，個々の文の構造を定式化した後，その構造に基づいて実際の文章を生成するとする手順をそれぞれモジュール化する手法が標準的である．

本研究では，前回の研究報告 [10] に引き続き，経済情報の記事を入力とした複数文書要約を，[6] に従い，談話構造の言語依存性に着目して試みる．また，時系列に基づいて複数文書から要約を整理することを試みる．

2 時系列データを入力とした月例経済報告の複数言語生成

まず，本研究の基礎として，単一の時系列データ入力からの複数言語生成を実現した．時系列データとしては，経済データを入力とした．本節では，まず Sripada[8] による時系列データからのデータ要約生成のための知識獲得について紹介し，次に実際の生成過程について説明する．

2.1 Sripada らによる時系列データの知識獲得

Sripada らはデータ要約を行うための SumTime プロジェクトを進めており，要約のための知識獲得として，以下の方針を採用している。

1. 要約のタスクモデルを決定する
2. タスクモデルに必要な知識の型を決定する
3. すべての必要な知識の型を詳細に決定する

SumTime プロジェクトではガスタービンデータや天気予報のようなデータ要約を対象としている。その文章を生成するために，専門化の知識を獲得したり，実際の天気予報の手で書かれたコーパス集合と時系列データを分析することを試みている．彼らの手法はドメイン独立の観察（ユーザに与える衝撃やデータの信頼性）とドメイン依存の観察（語彙の変化，内容の変化，非数値要因の分割）などに基づいている．この手法と同様に，実際の例文と同様の文章を XML 形式に基づいて格納した時系列データからデータ変換を行うことにより，複数言語生成を実現した．

2.2 時系列データベースからの月例経済報告生成

本研究では，まず，時系列数値データベースからの複数言語生成を実現した．入力としては，経済時系列データベース日経 NEEDS を使用し，内閣府発行の月例経済報告を日本語と英語を生成した．入力データは各月ごとに二，三ヶ月前のデータを項目ごとに獲得することになるが，時系列ごとに XML 形式で整理することで，問い合わせを一部変更するだけで各月ごとのデータを抽出することが可能となる．生成結果のうち，英語の月例経済報告の一部を付録 A に示す．この例は 2000 年 10 月の月例経済報告を模したものであるが，前年同月比や前月比など，一年前や二ヶ月前と三ヶ月前のデータを組み合わせて計算したり，もともとそのようなデータがある場合には，そのまま使用することで，必要な情報を XML-DB の問い合わせ言語を使用することで獲得している。また，値の正負に応じて，文脈に応じて “増減”や “縮小・拡大”，“上昇・下降”などの語彙の選択を行う．英文の生成にあたっては，語順や一文の構成ならびに語彙選択の基準が日本語と大幅に異なるため，文の構成に関するマイクロプランニングの段階ではもちろんのこと，文章全体のプランニングの段階でもやや異なる処理を行うことが望ましいとも考えられる．

本研究では，この結果に基づいて，テキストデータを入力とした複数言語要約の時系列データに基づいた実現に必要な技術について検討し，その技術に基づいて要約を実現する仕組みについて提案する．次節では，時系列データに基づく知識獲得に必要な技術について検討する．

3　時系列データに基づく知識獲得

前節の研究より，時系列に基づいてデータを整理することにより，時期に応じた文章を整理して提示することができることを示した．本研究では，時系列に応じてテキストデータを整理することにより，Sripada[8]の考えを応用して，複数文書要約を実現する．本節では，複数文書要約のアプローチについて紹介した後，本研究の関連研究として，Webページ群の時系列データからのトピック検出[1]およびWebマイニング[2]について紹介する。最後に，本研究で対象とする要約について説明する．

3.1　複数文書要約に向けてのアプローチ

複数文書要約は，テンプレートに基づく情報抽出[7]や，ソース文書群をクラスタリングすることにより，クラスタの代表要素を選択することで重複文書情報の除去する技術[9]に基づいて実現される．最近のものでは，[6]があり，ランク付けや選択のためのスコア付けに対する詳細な式を提案している。テキストの分類ならびに構成を行うためには，主題ごとに文書群から話題を構成する必要がある．

3.2　トピック検出と追跡技術

トピック検出と追跡技術については,（１）H-MM，情報検索や機械学習に基づくセグメント分割技術,（２）クラスタリング手法に基づくトピックの検出（３）古典的な情報検索のフィルタリング技術に基づくドメイン固有の追跡技術に基づいている．ある一つの話題についての重複した情報を避けるために，トピックごとにテキストをまとめる技術は，大量のソースからの複数文書要約においては非常に重要な技術である．ただし，新聞記事，特に，定期的に同じような話題が出てくる株式情報のような記事については，その時期がいつであるかといった情報が併せて重要となる。本研究では，類似性に基づいたトピック検出技術に時系列のデータ基準を取り入れることにより，類似性により本来異なる情報が要約から落ちることを避けることを試みる．

3.3　Webコンテンツマイニング

時系列データからの特徴的なデータ抽出技術としてはデータマイニングが良く知られているが，Webコンテンツマイニング技術[2]も同様のアプローチが期待される．トピック検出ならびに追跡の技術でテキストデータを整理した後，ドメイン独立に特徴的なデータを抽出することは，要約生成において重要な技術であるが，そのためにはテキストの内容のカテゴリーもしくは型を分類しておくこと必要となる．また，最新の情報が重要視されるかなど，目的に応じた重要度の計算方法も必要となる．

3.4　本研究で対象とする要約

本研究では，時系列ごとの類似した文書群の抽出という目的のために，相場報告を対象として要約を行う．具体的には，日本経済金融新聞の「相場を読む」と「今週の相場」を入力文書として採用した．これらの入力を上記の研究にしたがって解析し，一ヶ月ごとに，対応する相場の状況の要約を作成した．相場には「円相場」「債券市場」「金利」などの分類が可能であり，段落単位で区別して抽出する．また，中の文章は「原因となるトピック」「実際の数値データ」「上昇・下降などの状態」「時期」「出典データ名」「判断」などでXML形式でタグ付けする.「原因となるトピック」は,「目立っており」などのキーワードに応じて抽出される.「判断」は,「割安感」のような，直接的な表現を抽出する．

4　談話構造の言語依存性

要約を行うためには，談話構造の取り扱いが重要となる．ただし，談話構造は言語に依存して異なるとする報告[4]がある．本節では，複数言語生成のために必要なテキストプランナの役割[5]について紹介し，本研究との関連を示す．

4.1 複数言語要約に必要なテキストプランナの役割

複数言語生成において，談話構造や修辞関係は，言語に応じて異なり，パラレルコーパスから談話木を獲得することにより，言語依存の談話構造を獲得することができる．テキストプランナの役割としては，このような修辞関係よりの抽象的な構造と内容の決定を行い，言語独立な談話構造を設定し，各言語ごとに談話木を書き換えるアプローチも有効であるが，要約のようなテキストデータを入力とした場合には，[3, pp.156] にあるように，修辞関係の決定も含むため，談話構造は言語依存となる．

4.2 相場報告の言語依存談話構造

本研究の談話構造はまず，言語独立に円相場，債券市場，金利などの分類に基づいて入力文書を対応する要素を抽出して構成する．次に，その中で文書を時間順に並べる。また，tf/idfの値に基づいて，特殊性の高い文章の重要度を高くすることで，テキストデータ間の順序付けを行う．

以上の処理のあとに，言語依存の修辞構造プランニングを実現する．英語で日本語と異なるのは，月例経済報告生成のときと同様に，各要素の語順と，一文で表現する単位が異なる．二つの節を一つの文として構成する際に，日本語であれば一つの句である単位を英語では従属節として表現する必要がある場合，一文として適切な構成単位が異なる．本研究では，各言語ごとに句仕様として表現する単位を異なるものとして取り扱うことで，文書プランニングの段階で構造化の単位を出力言語に応じて変化させる．

5 複数言語要約の実現:Ruby&XML

数値データベースからの言語生成技術は，Rubyと XML を使用することで実現した．Ruby からのXMLの使用については，文書プランニングと文プランニングの段階でXML::SAX2+XML Parserを使用し，表層実現モジュールはXSLTプロセッサSablotron を使用して実現した．

6 おわりに

本研究では，相場に関する情報を一例として，時系列に基づいてテキストデータを整理することにより，複数文書から要約を生成することを試みた．また，整理したデータから談話に依存した修辞構造に基づいて談話構造をプランニングすることにより，複数言語要約生成へのテキストプランニングを言語依存に処理することにより実現する手法について提案を行った．

現在の段階の問題点としては，あらかじめ用意していないパタンについての情報が要約から抜けてしまっており，全文の情報が使用されていないことがあり，文抽出技術に基づく要約と比べて評価を取ることにより，本アプローチの有効性の評価を進めていく必要がある．

謝辞

本実験で使用したコーパスは,日本産業新聞・日本金融新聞2000年度版を使用させていただいた。使用に関してご尽力された方々に深く感謝します。また 日経NEEDSの内容につきましてご紹介いただきました日経メディアマーケッティング株式会社の北村雅人様に感謝いたします。

参考文献

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, February 1998.

[2] R. Kosala and H. Blockeel. Web mining research: A survey. In *Proc. of the 6th Int. Conf. on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)*, volume 2, pages 1–15, Boston, MA USA, July 2000.

[3] I. Mani. *Automatic Summarization*, volume 3 of *Natural Language Processing*. John Benjamins, Amsterdam, Philadelphia, first edition, 2001.

[4] D. Marcu, L. Carlson, and M. Watanabe. The automatic translation of discourse structures. In *ANLP-NAACL 2000*, Seattle, WA USA, May 2000.

[5] D. Marcu, L. Carlson, and M. Watanabe. An empirical study in multilingual natural language generation: What should a text plannner do? In *the 1st Int. Conf. on Natural Language Generation (INLG'2000)*, Mitszpe Ramon, Israel, June 2000.

[6] D. Marcu and L. Gerher. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proc. of the NAACL-2001 Workshop on Automatic Summarization*, Pittsburgh, PA, June 2001.

[7] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In *the 18th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 74–82, Seattle, WA USA, July 1995.

[8] S. G. Sripada, E. Reiter, J. Hunter, J. Yu, and I. P. Davy. Modelling the task of summarising time series data using ka techniques. In *Proc. of ES2001*, 2001.

[9] G. C. Stein, T. Strzalkowski, and G. B. Wise. Summarizing multiple documents using text extraction and interactive clustering. In *Pacific Association for Computational Linguistics (PACLING-1999)*, 1999.

[10] 関 洋平, 原田 賢一, and 野村 直之. Rubyによる複数資源要約システムの実現. In **情報処理学会情報学基礎・デジタルドキュメント合同研究会** *FI66-DD32-7*, pages 47–54, March 2002.

A 月例経済報告の生成結果（英語）

1. Domestic Demands

Personal Consumption

Living expenditures（whole）for July decreased 2.6 % compared to the same period last year, and for August a 4.1 % decrease compared to the same period last year.

When you look at the change classified by household spending, there was a 2.9 % decrease compared to the same period last year for working people in August.

The consumption level for August decreased 3.09 % compared to the same period last year.

The consumption level for working people in August decreased 2.09 % compared to the same period last year.

Wages

Income for August decreased 1.19 % compared to the same period last year for companies employing 30 or more people.

Additional allowances for August decreased 5.46 % compared to the same period last year for companies employing 30 or more people.

Real wages for August decreased 2.12 % compared to the same period last year for companies employing 30 or more people.

Housing Construction

The number of housing starts（seasonally adjusted rate） for July decreased 2.44 % compared to the last month, and a 0.53 % decrease compared to the same period last year. The number of housing starts（seasonally adjusted rate） for August decreased 0.11 % compared to the same period last year.

The floor space of new houses for August decreased 0.93 % compared to the last month, and a 2.30 %

decrease compared to the same period last year.