E-6　固有表現を利用した大量文書の時系列ブラウジング法

A Chronological Browsing Method of Language Corpus Using Named Entity

國領 弘治†　佐々木 裕†　前田 英作†
Kouji Kokuryou　Yutaka Sasaki　Eisaku Maeda

1．はじめに

新聞記事に代表される「時間情報をもった大量文書」は、特定のトピックに関する情報を時間軸に沿って得ることができ、情報源としての価値が高い。しかしながら、こうした情報源を有効活用するためには、大量の記事から特定のトピックに関する記事を抜き出し、かつ抜き出された記事から必要な情報のみを効果的に提示することが不可欠である。

本稿では、こうした手法を「時系列ブラウジング法」と呼び、その解決策として固有表現を利用した時系列ブラウジング法を提案する。過去からの蓄積情報である新聞記事には、断続的に現れる話題（スペースシャトルや APEC 首脳会議など）があり、これらの話題について検索した場合、大量の関連記事が検索され、そこから必要な情報を適切に抜き出すことは容易ではない。

11 年分の新聞記事を検索し、検索結果N件の中から解答候補を抽出する“日本語質問応答システム（SAIQA）[1]”においても同様で、大量記事から得られる情報を効率的に、時系列ブラウジングできれば、様々な応用が考えられる。

本稿では、新聞記事が持つ時間情報に沿って特定のトピックを可視化し、固有表現を用いて記事のエッセンスを提示することを可能とする「時系列ブラウジング法」を、日本語質問応答システム(SAIQA)に実装し、その効果を確認した。

2．条件設定

時系列ブラウジング法の入出力条件は以下の通り

入力　：過去数年間に周期的にピークが現れる話題語
出力　：話題の周期を確認する為のグラフ及びピーク時の特徴キーワード
データ：91 年から 99 年までの毎日新聞の電子テキストを使用。

3．時系列ブラウジング法の概要

検索語により９年分の速報記事を検索、検索結果を時間情報で分類した後、集合の規模に応じたグラフと、集合中の特徴キーワードを提示する。

3．1 記事分類

新聞記事を大きく分類すると、予定記事、速報記事、解説記事に分けられる。

予定記事：数日から数ヶ月先の話題について述べたもので、正確な時間情報が得られない。また予定が変更になる場合もある。
・速報記事：話題発生後すぐに書かれたもので、数日の誤差はあるが発行日を時間情報として使える。
・解説記事：数日から数ヶ月前、場合によっては数年前の話題について述べたものもあり、時間情報が記述されていない場合が多い。

このうち時間情報として最も正確で入手しやすい発行日が利用できる“速報記事”を用いる。

記事の分類方法には、頻出後の共起関係から文書構造の特長を得るものや[2]、照応現象や文の連接関係から文書構造をモデル化するものなど[3]、文書全体を捉えた分類もあるが、“速報記事“の特徴である 1 文目の文末表現に着目し分類した。

速報記事の文末表現

動詞(自律)＋助動詞	(例) ～が明らかになった
動詞(接尾)＋助動詞	(例) ～が発見された
サ変名詞	(例) ～を決定

3．2 グラフによる可視化

・速報記事に付与された発行日のうち、年月のみを時間情報として用い、検索結果を月単位の集合に分類する。

・集合中の記事数を単位とした棒グラフを提示する。

・速報記事の総数と出現月数から閾値を求め、閾値以上の記事数であれば、特徴キーワードを提示する。

3．3 特徴キーワード

・周期的に現れる話題を特徴付けるキーワードとして、人名(PER)、人工物名(ART)、場所(LOC)、組織(ORG)などの固有表現[4]を用いる。

・出現記事数や出現頻度などの情報からその集合を特徴付ける固有表現を数個決定し、重要度順に左から提示する。この際、上位２位とそれ以外とを色分けするなど、視覚上の工夫も施した。

† 日本電信電話株式会社
ＮＴＴ コミュニケーション科学基礎研究所
NTT Communication Science Laboratories,
NTT Corporation

・また、”人名”などは出現頻度が低くても重要なキーワードであることから、固有表現のタイプ毎に提示することとし、重要な固有表現ほど上位に配置した。

3．4 特徴キーワードの補助情報

固有表現を提示しただけでは特徴が掴めない場合がある。そこで、人名(PER)に対する“役職情報”など、特徴キーワードを補助する情報を付加することにした。現在は人名(PER)の補助情報のみ提示。

4．実行例

話題語を「スペースシャトル」として時系列ブラウジングを行った時の実行例を図1に示す。なお、* は記事数を表している。

92年9月、94年7月、98年11月にピークが現れ、日本人宇宙飛行士の名前とスペースシャトルの愛称が、特徴キーワードとして得られた。91年から99年の9年間に5つのピークが現れるが、表1に示す2つの固有表現で周期特定が可能になることが分かった。

図2は話題語を「APEC首脳会議」とした例である。

毎年 11 月にピークが現れ、開催地と開催国および出席した各国首脳の名前が、特徴キーワードとして得られた。

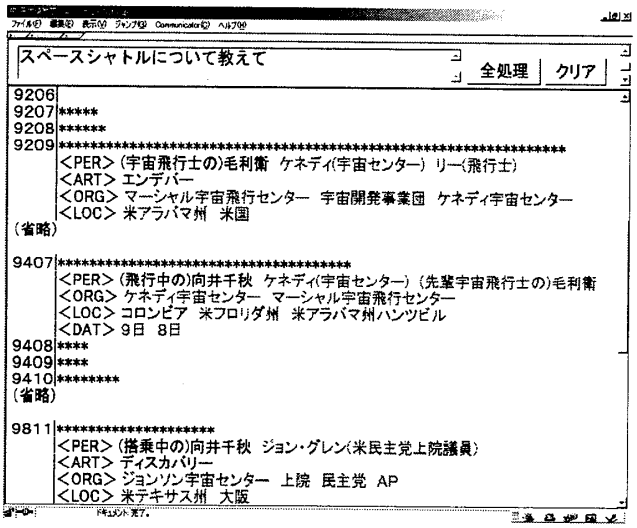

図1：時系列ブラウジングの出力(スペースシャトル)

表1：話題のピークと固有表現の例

話題のピーク	人名(PER)	人工物名(ART)
92年9月	毛利 衛	エンデバー
94年7月	向井 千秋	コロンビア
96年1月	若田 光一	エンデバー
97年11月	土井 隆雄	コロンビア
98年11月	向井 千秋	ディスカバリー

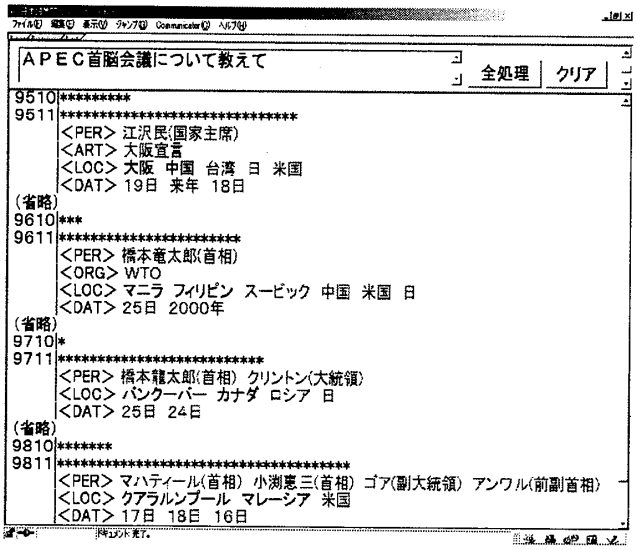

図2：時系列ブラウジングの出力(APEC首脳会議)

5．関連研究

ユーザ支援を目的とした技術には“適合性フィードバック”や“対話に基づく検索”などがある。[5]

適合性フィードバックとは、システムが提示した検索結果をユーザが「適合」「不適合」に分類することで、システム側からのフィードバック情報(ベクトル空間モデルの場合、検索語の重みや追加する検索語など)を得るものである。

また、対話に基づく検索である“THOMAS system”の場合は、ユーザが1つないし2つの検索語を入力、システムが検索結果と文書の概念を表す検索語を提示、ユーザが検索結果及び検索語の要否判定を行う、検索語の提示と要否判定の繰り返しでユーザ要求を具体化していくものである。

いずれも検索結果から得られた情報をもとに、検索語を充実させることでユーザを支援する、という点で本研究の目的に共通するものである。

6．おわりに

本稿では、時間情報を持った大量文書の問題点を補う、時系列ブラウジング法について述べた。今後は、オープンドメインな情報抽出の研究などと関連付けていきたい。

参考文献

[1]佐々木 裕，磯崎 秀樹，平 博順，平尾 努，賀沢 秀人，鈴木 潤，國領 弘治，前田 英作：“SAIQA：大量文書に基づく質問応答システム”，自然言語処理研究会，2001-NL-145，2001

[2]加藤優，松尾豊，石塚満：“文書の構造に基づくクラスタリング”，第2回AI若手の集い MYCOM2001(2001)

[3]福本淳一，安原宏：“日本語文章の構造化解析“，　自然言語処理研究会，1991-NL-85，1991

[4]磯崎秀樹：“ＳＶＭに基づく固有表現抽出の高速化”，自然言語処理研究会，2002-NL-149，2002

[5]徳永健伸：“言語と計算-5　情報検索と言語処理”，東京大学出版会，pp.154-169