3H9-5

情報編纂のための動向情報タグ付けシステムの構築

Construction of Trend Information Tagging System for Information Compilation

上西康広[*1] 今岡裕貴[*1] 桝井文人[*1] 河合敦夫[*1] 井須尚紀[*1]
Yasuhiro UENISHI Hirotaka IMAOKA Fumito MASUI Atsuo KAWAI Naoki ISU

[*1] 三重大学大学院工学研究科情報工学専攻
Department of Information Engineering, Graduate School of Engineering, Mie University

We aim at construction of a general-purpose trend information tagging system. We are advancing in study about function to reason with the trend information that is not stated clearly in text by the use of relative expression and function to extract trend information in text by application of Template Relation(TR).We are advancing in study of trend information reasoning function that used relative expression and confirmed that F-measure shows about 0.8 performance till now. Now, to realize trend information extraction mechanism by application of TR, we push forward investigation about to a characteristic and tendency of trend information among Named Entity in newspaper article. In this paper, we discuss outline about trend information tagging system and trend information findings among Named Entity in newspaper article.

1.はじめに

IT化の促進に伴い，様々な情報が大量に電子化され，増加し続けている．このような玉石混合の情報洪水の中から有用な情報のみを効率良く選択して利用することは現代社会において非常に重要であるが，同時に非常に難しい問題でもある．

このような玉石混合状態から玉を取り出すためには，ユーザの関心に応じた柔軟な情報編纂技術が必要である．情報編纂指向[1]の研究テーマとしては，最近，動向情報の要約・可視化[2]が注目されている．動向情報とは，ある商品の売り上げ，ある業界のメーカー別業績，内閣の支持率など，時々刻々と変化する数値に関連した情報である．このような場合，グラフや図，地図などを用いて視覚化した方が直感的に理解し易い場合が多い．そのため，テキスト中に記述された動向情報を効率的に把握する支援技術として，テキスト情報を解析して動向情報を自動抽出し，さらに，ユーザの関心に応じて最適な視覚情報として再構成する技術は非常に有効である．

我々は，ユーザの関心を把握し，それに適した形式で動向情報を提供するシステムの開発を計画している． 本システムは，ユーザが意識することなく，ユーザの関心を把握し，その関心に基づいて，ユーザにとって有用であろう動向情報を逐次提示する．システムは，以下に示す三つのサブシステムから構成される．

（1）ユーザの関心を把握するサブシステム
（2）動向情報を抽出するサブシステム
（3）適切な形で動向情報を提示するサブシステム

本論文では，ユーザの関心に追従する動向情報提示システムの概要と，その中核システムとして位置付けられる「動向情報抽出システム」について説明する．

2.ユーザの関心に追従する動向情報提示システム

2.1 ユーザの関心を把握するサブシステム

ユーザの関心を把握し，関心に応じたテキスト情報を出力する．このとき，語や文，パッセージや文章，段落といった多様な言語情報を利用してユーザの関心を把握する．語や文，パッセージのレベルであれば，情報検索や質問応答など，ユーザが明示的に関心を表現できる場合が想定できる．これに対して，文章や段落となると，ユーザが意識的に提示する関心というよりも，その時点においてユーザが意図的ではなく，漠然と関心を示しているテキスト情報である場面を想定する方が自然である．さらにユーザの関心が動的に変遷していく状況も考えられる．

2.2 動向情報抽出サブシステム

（2）では，（1）で限定されたテキスト情報を対象として，必要な動向情報を抽出する．抽出した情報は，タグ付けされた状態，あるいはデータベースやテンプレートなど形式化された状態で出力される．動向情報抽出における基本的技術として，情報抽出技術[3]を応用して用いる．

ここで，処理対象となるテキスト情報はユーザの関心に対応する．したがって，処理対象テキストは，複数のドメインやトピックに跨る可能性がある．例えば，「アサヒビール」に関するテキスト情報が対象となっている場合を考えると，「株価」や「ビール出荷数量」，「業務提携」など，様々なトピックやドメインを含む可能性がある．これは，ドメイン依存を前提としていた従来の情報抽出の範疇を越える．よって，トピックやドメインに依存しない情報抽出の拡張アプローチが必要となる．

2.3 動向情報提示サブシステム

（3）では，（2）で抽出した動向情報を編纂し，適切な形で提示する．適切な形で提示するためには，ユーザの関心，情報の種類や特徴，形式などを考慮する必要がある．例えば，株価の変動やビールの出荷数量などはグラフで示すべきである．一方，台風の進路や地震の震度などは地図上に示すべきである．

さらに，ユーザの関心に合った見せ方をする必要がある．例えば，「2000年のビールの出荷数量」についての情報があったとする．このとき，ユーザの関心が「企業別に見たい」という場合は積み上げ棒グラフで示すことが考えられる．しかし，「月ごとに見たい」という場合は折れ線グラフなどが考えられる．また，同じ情報でもユーザの関心により，提示する順序を変える必要もある．

連絡先：三重大学大学院工学研究科情報工学専攻
〒514-8507 三重県津市栗真町屋町 1577
{uenishi, imaoka, masui, kawai, isu}@ai.info.mie-u.ac.jp

3.動向情報タグ付けサブシステム

前章で述べた三つのサブシステムのうち，(2)動向情報を抽出するサブシステムは，処理全体の中心的な役割を果たす．よって我々は，本サブシステム構築を第一の目標とする．

本サブシステムは，ユーザの関心に対応するテキスト情報入力とし，動向情報を抽出する．このとき，抽出した情報の出力形式には，タグ付きテキスト，データベース，テンプレートなど複数のバリエーションが考えられる．後処理における復元可能性を考慮し，デフォルト出力として図1のようなタグ付けを行うる．

```
<company type="beer company TR="ラガー" TR_type="product_of">キリン</company>
```

図 1:タグ付けの例

以下，本サブシステムにおいて抽出対象となる情報，および抽出処理における技術的課題について，MUCで提案された情報抽出タスクとの対比を意識しながら述べる．

3.1 動向情報として重要な要素の抽出

動向情報を抽出するためには，対象テキスト情報から動向情報において重要な役割を占める語句情報を特定・抽出する必要がある．それらは多くの場合，「アサヒビール」のような組織名に加え人名や地名，さらに「1998年」のような時間表現，数量表現であり，固有表現(NE)に相当する．この点で，ここで述べる語句情報の特定・抽出は，固有表現抽出タスクと同等である．しかしながら，「ビール出荷量」や「株価」，「著作権侵害」など，固有表現の範疇には収まらない名詞句も存在する．これらの語句をも抽出対象とするためには，固有表現抽出における抽出対象を拡張した処理を考える必要がある．

また，ユーザの関心やテキスト中のトピックによっては，「アサヒビール」のタイプは「組織名」であったり，「酒造メーカー」であったりする．これらの判断は，従来のTemplate Elementタスクを拡張することで対応する必要がある．

3.2 動向情報として重要な要素間の関係の特定

前節で説明した抽出された重要要素は相互に特定の関連を持っている可能性が高い．例えば，「アサヒビール」と「スーパードライ」という要素の間には「製造元と製品」という関連があったり，「アサヒビール」と「2345円」という要素の間には「株式銘柄と株価」といった関連がある．

このように動向情報の重要要素間の関係を特定する処理においては，固有表現間の関係を特定するTemplate Relation (TR)タスクが応用できそうである．しかし，本論文において処理対象となるテキスト情報はユーザの関心に対応する．したがって，処理対象テキストは，複数のドメインやトピックに跨る可能性があるため，ドメイン依存を前提としていたTRタスクの範疇を越えてしまう．よって，トピックやドメインに依存しないTRの拡張アプローチが必要となる．

このようなドメイン依存の問題を解消するためのアプローチとしては，以下のような方法が考えられる．ひとつは，与えられた文書集合に存在するTRをシステムが抽出可能なTRを全て抽出する場合である．この場合，どの情報をユーザに見せるかという判断は，(3)抽出した情報を適切な形で提示する段階で実施することになる．

もうひとつは，ユーザの関心とユーザの関心にあった文書集合から，その都度抽出するTRを選択する場合である．この場合，ユーザの関心にあった文書に対して，属するトピックを決めてしまう．その際に問題となるのが，与えられた文書集合が複数のトピックに属する可能性があるということである．複数のトピックへの対応については(A)対象テキスト情報が複数のトピックに属するか否か，(B)対象テキスト情報を分割するか否か，という二つの観点が考えられる．抽出過程は，この二つの観点の組合せ方によって複数実施できる．

たとえば，もっとも単純な例として，対象テキスト情報が属するトピックがひとつに決まり，対象テキスト情報もひとつである場合が考えられる．この場合はトピックが一意に決定できるため，単純なTRタスクを実施すればよい．

一方，対象テキスト情報が複数のトピックに属するが，対象テキスト情報はひとつである場合を考えると，同一テキスト情報に対して異なるトピックにおけるTRタスクを複数回実施する処理を考える必要がある．さらに，対象テキスト情報が複数トピックに属し，かつ対象テキスト情報も複数テキスト情報に分割できる場合も考えられる．この場合，最初に述べたTRタスクが複数実施される状況となる．

上記で述べた拡張TRタスクは，あらかじめ複数のトピックにおけるTRのテンプレートが特定できていれば，いずれも単純なTRタスクの組合せとして実現できる．これについて，実現可能性を見極めるために，MuSTコーパス[2]を対象として，コーパス中で定義されている幾つかのトピックにおけるTRの出現分布を分析する予備調査を実施した．その結果，例えばビールトピックでは「製品と出荷量」が 15%，自動車トピックでは「組織と組織」が 38%というように，トピック毎に主要なTRを特定することが可能であることがわかった．

上記以外にも，より複雑な拡張TRタスクが考えられる．また，抽出された複数のTR間にも関係が発生し，それらが動向情報を構成する可能性もある．このような複雑な抽出処理についても議論する必要がある．

3.3 照応関係の特定

抽出の対象となっている情報が複数の文で表現され，最初の文にあった名前が二文目以降では，代名詞，省略形，ゼロ代名詞などで表されていることがある．例えば，最初の文で「アサヒビール」と書かれていたものが，次に出現する文では「同社」，「アサヒ」と書かれている場合がある．

また，時間表現についても，同様の問題がある．これは，例えば「2000年9月16日」という時間表現は，文書中では，「16日」，「昨日」などと様々な表現がなされている．適切に情報を抽出するためには，これらの表現が何を指しているのかという照応関係を特定する必要がある．

4.おわりに

本論文では，ユーザの関心に追従する動向情報提示システムの概要と，その中核システムとして位置付けられる「動向情報抽出システム」について述べた．現在までに，相対表現を利用した，テキスト中に明示的に現れない動向情報を推論する方法を提案し，F値で約 0.8 の性能が得られることを確認した．今後は，従来の情報抽出技術を拡張させた動向情報抽出について研究を進める予定である．

参考文献

[1] 加藤恒昭，松下光範 ： 情報編纂（Information Compilation）の基盤技術，2006 年度人工知能学会全国大会（第 20 回）論文集，2006.

[2] 松下光範，加藤恒昭 ： 動向情報に基づく情報可視化の基礎検討，2005 年度人工知能学会全国大会（第 19 回）論文集，2005.

[3] Nancy Chinchor : MUC-7 Information Extraction Task Definition version 4.2, 1998.

[4] 桝井文人，鈴木伸哉，福本淳一 ： テキスト処理のための固有表現抽出ツール NExT の開発，第 8 回言語処理学会論文集，pp.176-179，2002.