相対表現に基づく動向情報抽出システムの改良とその評価

An Improvement in the Relative Expression-based Trend Information Extraction System

桝井 文人[*1] 上西 康広[*2] 松葉 達明[*2] 河合 敦夫[*2] 井須 尚紀[*2]
Fumito MASUI Yasuhiro Uenishi Tatsuaki MATSUBA Atsuo KAWAI Naoki ISU

[*1]北見工業大学 情報システム工学科
Department of Computer Science, Kitami Institute of Technology

[*2]三重大学 大学院工学研究科
Graduated School of Engineering, Mie University

This paper describes a system extracting trend information. We applied the method of implicit trend information extraction utilizing relative expression such as “0.1%増 (grew 0.1%”, “前年 (previous year)” to the system..Relative differences and numerical changes in trend information can be signified by relative expressions. The system extracts elements of four types by patters-based rules considering the relative expression. The extracted element is compared with the query word by identifying and expanded the synonym of the elements utilizing an EDR dictionary and some synonym databases. An experiment was conducted withe the enlarged MuST corpus. The results showed precision of 0.638 and recall of 0.287 totally, and our system overcome the former system.

1. はじめに

膨大な文書データからユーザに有用な情報のみを効率よく抽出・整理して提示する技術として，動向情報の要約・可視化が注目され，研究用コーパスも公開されている [Kato&Matsushita 08]. 動向情報抽出とは「06 年からのゲーム業界はどうなっているか」「今年のガソリン価格どう推移しているか」といったユーザの関心に対して基礎的情報を提供する技術である．

動向情報の一部として出現する統計量や日付表現には「前年比 10%増」のような数値の相対的な差異や，数値の変動を示すものがあり，相対表現と呼ばれる．相対表現と明示的な動向情報を組み合わせることによって，テキスト中に明示されていない情報を推論することができる．テキスト中の動向情報は基本的に，{ *name* (統計量名) , *par* (パラメータ) , *date* (日付) , *val* (統計量)} といった要素の組み合せ（以降，四つ組とよぶ）で示すことができる．例えば，「**2007** 年のアサヒのビール出荷量は前年比 **0.1%**増の **1** 億 **8824** 万ケースとなった。」という文からは，{ ビール出荷量, アサヒ, **2007** 年, **1** 億 **8824** 万ケース } という明示的 (explicit) な四つ組が得られる．ここで「前年比 0.1%増」という相対表現があり，上記の四つ組に適用した場合，{ ビール出荷量, アサヒ, **2006** 年, **1** 億 **8800** 万ケース } という，テキスト中に明示されない（implicit な）四つ組を導出できる．

我々は上記のような効果を使った動向情報抽出を実現するための研究を進めて来た．入力クエリに対応した相対表現の生成パターンを抽象化し，明示的な動向基本情報を抽出する機構を構築した [今岡ら 2006]．対象文字列を headword と specifier の二つに分割してクエリ解析を行うことで，動向情報を構成する四つ組を効率よく抽出する手法について考えてきた [Uenishi et. al 09] [上西ら 09]．

本論文では，クエリ分割によって生成した形態素系列の比較を工夫することで，クエリバリエーションの推論を精緻化する．

さらに，従来の研究では，各手法の評価は MuST T2N subtask のテストコレクションを用いたものであったが，規模が不十分であった．そこで本論文では，実験規模を拡大して評価実験を行った．

連絡先: 桝井文人，北見工業大学 情報システム工学科，北海道北見市公園町 165, Email: f-masui@mail.kitami-it.ac.jp

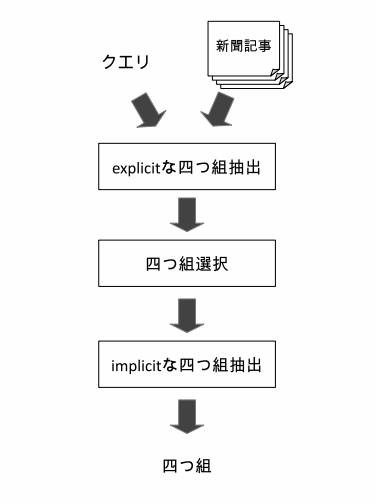

図 1: 提案手法の処理概要

以下，2 章で提案手法の概要を説明し，3 章では評価実験について述べ，4 章で実験結果について考察する．

2. 提案手法の概要

本章では，提案手法の概要について述べる．本手法は，相対表現に基づき，与えられた統計量名（クエリ）と関連する explicit な四つ組をタグ付き文書から抽出し，抽出時の相対表現に対応した推論規則によって implicit な四つ組を生成する．提案手法は図 1 のように四つのモジュールから構成される．

クエリ解析・文書検索モジュールは以下の様に処理を行う．まず，辞書知識を用いて headword と spefier という二つの構成単位に分割する．構成単位の言換えによってクエリバリエーション候補を生成する．言換えには，EDR 概念辞書，構成単位とその言換え知識を記述した headword 辞書および specifier 辞書を用いた．

次に，クエリバリエーション候補を形態素解析し，名詞・未

知語・接頭詞による形態素系列に変換する．生成できる全ての形態素系列間の比較を行うことでクエリバリエーションを推論する．形態そ系列間の比較において，[上西ら 09] では包含関係を同一性条件に含めていたが，提案手法では形態素系列の要素数が一致する場合のみを同一性条件とした．

最後に，生成されたクエリバリエーションを文書検索して適合フィードバックを行う．ここで有意な出現数を持つクエリバリエーションを認定する．

四つ組の選択モジュールでは，クエリ解析で獲得したクエリリストの各エントリと四つ組を構成する name 要素と par 要素を用いる．

まず，獲得された四つ組集合から一組を取り出す．取り出した四つ組の name 要素と par 要素を形態素解析し，名詞・未知語・接頭辞による形態素系列 (N, P) に変換する．得られた二つの系列を $P + N$ のように結合して新たな系列 Q を生成する．クエリバリエーションから形態素系列 q_i を取り出し，系列 Q と比較を行うことで関連性判定を行う．系列 q_i と系列 Q が完全一致する場合に関連性ありと判定する．

上記の処理を全ての四つ組に対して実施することで，適切な四つ組のみが選択され，各四つ組中の要素を用いて対応した推論規則を適用することで implicit な四つ組が生成される．

3. 実験と評価

本章では， 前章までに述べた提案手法の有効性を検証するために評価実験を行う．4 つ組選択手法を 4 章で述べた相対表現を利用した動向情報抽出に適用しシステムを構築した．

ベースラインとして [今岡ら 06](baseline)，対象実験として [Uenishi et. al. 09]，[上西ら 09] をそれぞれ実装し，提案手法と比較した．

入力データとして， NTCIR MuST T2N subtask [Kato&Matsushita 2008] で用いたクエリ 19 個，2006 年度の MuST コーパス中に現れる相対表現のうち出現頻度が高いトピックに含まれるクエリ 8 個，計 27 個のクエリを用いた．

抽出元コーパスとして，上記タスクで用いた 1998 年〜2001 年のタグ付き毎日新聞 100 記事 (MuST コーパス) と，毎日新聞 (2000 年から 2003 年) の 43 記事を使用した．追加の 43 記事に対しては人手で MuST コーパスと同様のタグを付与した．

本研究では各システムの差異は相対表現に関連した四つ組の抽出処理であるため，explicit な四つ組の抽出性能を評価対象とした．それぞれの評価値の定義を以下示す．

$$\text{適合率 } P = \frac{\text{正しく選択できた四つ組の数}}{\text{システムが選択した四つ組の数}} \quad (1)$$

$$\text{再現率 } R = \frac{\text{正しく選択できた四つ組の数}}{\text{選択すべき四つ組の数}} \quad (2)$$

$$F \text{ 値} = \frac{1}{\frac{1}{P}\alpha + \frac{1}{R}(1-\alpha)}{}^{*1} \quad (3)$$

評価は MuST T2N subtask で配布された正解データを使用して評価を行った．追加の 43 記事については， MuST T2N subtask の正解データに従って， 正解データを人手で作成し，評価を行った．表 1 に評価結果を示す．

4. 考察

本章では，提案手法の評価結果について考察する．全体的にみると，適合率と F 値で提案手法が最も高い性能を示しており，提案手法の基本的な有効性が確認できたといえる．

表 1: 従来手法との抽出性能の比較

		baseline	[Uenishi 09]	[上西ら 09]	提案手法
適合率	macro ave.	0.531	0.542	0.523	**0.638**
再現率	macro ave.	0.214	0.294	**0.315**	0.287
F 値	macro ave.	0.285	0.360	0.367	**0.379**

再現率については，提案手法は 0.287 と [上西ら 09] や [Uenishi et. al. 09] よりも低い値となってしまった．これは，関連性判定基準において，[上西ら 09] の手法が， クエリと完全一致に加えて表層的包含関係をも抽出対象としていることが主な理由である．例えば，クエリ「新設住宅着工戸数」に対して，「新設住宅着工の総戸数」などの関連性判定が成功例である．しかし，クエリ「受験率」に対して「公民の受験率」などを関連性ありと判定するケースが多く，再現率への寄与よりも適合率への悪影響の方が大きくなってしまっている．

表 2 に提案手法の抽出失敗原因の内訳を示す．

表 2: 正解を抽出できなかった原因の内訳

原因	割合 (%)
検索ミス	18
パターンマッチ	28
不足要素補完	20
同一性判定	34

最も割合が大きな同一性の判定誤りの原因は， 記事中でしばしば出現する「出荷台数」のような specifier が省略された表現や「携帯電話」のような headword が省略された表現に対応しきれなかったためである．適切な判定を行うには，機械学習などを用いて省略されている形態素を補完して統計量名を大量に生成する必要がある．

5. おわりに

本論文では，クエリ構成単位を辞書を用いて拡張し，クエリの同義バリエーションを作成することによって同一性判定の性能を向上させる方法を提案した．さらに，テストデータを作成して従来より実験規模を拡大して評価実験を行った．その結果，explicit な四つ組の抽出性能において従来よりも高い性能が確認でき，相対表現を利用した動向情報抽出への効果が大きいことがわかった．

謝辞

本研究は科研費（20500833）の助成を受けたものである．

参考文献

[Kato&Matsuthita 08] Kato, T., and Matsushita, N.: Overview of MuST at the NTCIR-7 Workshop – Challenges to Multimodal Summarization for Trend Information –, Proceedings of NTCIR-7 Workshop Meeting, pp. 475-488(2008)

[今岡ら 06] 今岡 裕貴, 桝井 文人, 河合 敦夫, 井須 尚紀: 動向情報抽出における相対表現の利用効果に関する考察, 日本知能情報ファジィ学会誌, 18 (5), 735-744(2006)

[Uenishi et.al 09] Uenishi, Y., Matsuba, T., Masui, F, Kawai, A. and Isu, N: Trend Information Extraction based on Relative Expression participated on MuST T2N Subtask, Proceedings of 7th NTCIR Workshop Meetings, pp.509-514(2008)

[上西ら 09] 上西康広, 松葉達明, 桝井文人, 河合敦夫, 井須尚紀: 相対表現に基づく動向情報抽出システムの構築, 言語処理学会第 14 回年次大会発表論文集,P1-10(2009)