1L−3

# 係り受け関係を利用した新製品紹介記事からの製品情報抽出 ～紹介製品数を区別しない抽出法～

赤松 順子　永井 秀利　中村 貞吾　野村 浩郷†

九州工業大学　情報工学部†

## 1 背景

近年さまざまな情報が電子化され，氾濫している情報を管理する技術として，大量の文章データの中から目的の情報のみを取り出してくる，情報抽出の技術が要求されている．

従来の研究では，定型性があり，抽出対象が明確である文書に対して，字面処理によって高速に情報の抽出を行うという方針の基，新製品紹介記事を対象に，抽出項目と周辺文字列から成るテンプレートと入力文章をマッチングするという手法で抽出を行なってきた．

その結果，1記事に1つの製品を紹介している記事に限定した実験[1]では高い精度の抽出に成功したが，複数の製品を紹介している記事は文章の構成が多様化し，抽出処理に加えて，製品項目間の対応付けが必要となるため，単純なマッチング処理による高精度の抽出は困難であった[2]．また，これまでは，1つの製品，複数の製品と対象を分けてシステムを構築していたため，これらを組み合わせたシステムでは，双方の処理に違いなどから，抽出精度，抽出時間ともに悪化すると考えられる．

そこで本論文では，複数製品の記事の精度向上を目的として行なっていた構文解析結果の一部を利用した抽出手法[2]を改良し，すべての新製品紹介記事からの情報抽出を行なうことを目的としている．

## 2 新製品紹介記事の形式と抽出項目

新製品紹介記事として，毎日新聞1991～1995年のデータから，新製品紹介記事とみなせる記事を人手で収集し，91，92，94，95年を分析用データとした．以下の記事を例にとって，抽出する項目を示す．

三洋電機はニューロ＆ファジーを採用した衣類乾燥機「さっ速ドライ」を９月１日、発売する。
３つの温度センサーで布質、量、含水量などを…
～中略～
…乾燥時間が従来より約３０％短縮されて約３０分で済み、電気代も３円程度安くなる。
乾燥容量４キロタイプ７万２５００円、４．５キロタイプ８万３５００円。

新聞記事例

| | 製品１ | 製品２ |
|---|---|---|
| 販売元 | 三洋電機 | |
| 発売日 | ９月１日 | |
| 製品種別 | 衣類乾燥機 | |
| 製品名 | 「さっ速ドライ」 | |
| 細分類 | ４キロタイプ | ４．５キロタイプ |
| 価格 | ８万３５００円 | ７万２５００円 |

抽出項目

## 3 抽出手法

例のように，記事の構成は”発売する”などの文末表現で製品の発売を明記する文，製品の特徴を説明する文，価格を示す文という定型性が見られた．抽出項目出現位置の分析から，必要な情報の大部分は，発売を表す動詞，その動詞に係るまたは並列関係にある動詞，価格の3種類に係る文節に出現した．そこで，抽出項目の係り先となる，製品発売を意味する表現を固定パターンとして定義しておき，構文解析器KNP[4]を用いて，固定パターン，価格の表記と係り受け関係にある文節を求め抽出を行なった．

主な固定パターンと，抽出項目が固定パターン1に係る格の形式のうち頻出するものを挙げる（“発売＊”は「発売する」，「発売した」などを表す）．なお，固定パターン2に係る場合の格の形式は，それぞれの意味に応じて分類し，それぞれに格を設定している．

【固定パターン】

| 固定パターン1(*1) | 固定パターン2(*2) |
|---|---|
| 発売＊，販売＊，開発＊<br>発表＊，売り＊，発表＊ | 追加＊，搭載＊，，開発＊<br>改良＊，装備＊，，採用＊<br>設定＊，＊チェンジ＊<br>リニューアル＊，輸入＊ |

*1 固定パターン1：文末表現

*2 固定パターン2：1に係る又は1と並列の表現

【抽出項目の格形式】

| | |
|---|---|
| 販売元 | 未格，ガ格 |
| 発売日 | カラ格，隣接，無格，ニ格 |
| 製品種別 | ヲ格，同格連体(*3) |
| 製品名 | ヲ格 |
| 細分類(*4) | デ格，ガ格 |
| 価格(*5) | — |

*3 製品名に係る場合

*4 価格に係る場合

*5 価格は表記パターンのマッチングで抽出

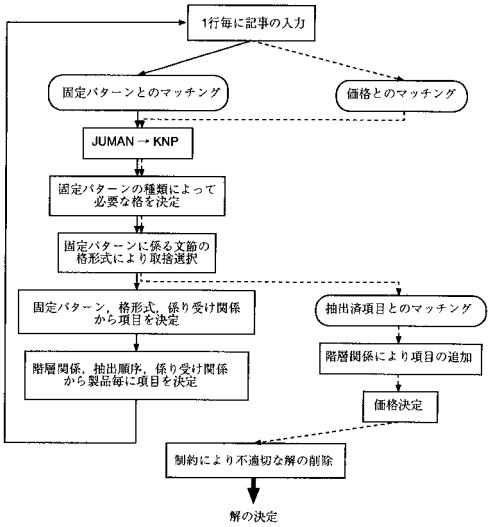
1記事中に1つの製品を紹介した記事では，固定パターンと価格との係り受けのみで解を決定できるが，複数の製品を紹介している場合は各製品毎に抽出項目を対応付ける必要がある．これには，1文中の抽出項目間の係り受けをまとめた情報と，複数製品紹介記事の分析[3]から得た，以下の項目間の階層関係を用いて対応付けを行なった．

| 販売元 | 発売日 | 製品種別 | 製品名 | 細分類 | 価格 |
|---|---|---|---|---|---|

上位←　→下位

項目の階層関係

構文解析を用いた情報抽出実験システムについて述べる．本システムの概要図を示す．




製品情報抽出システムの構造

## 4 実験

実験では，毎日新聞 1991 年から 1995 年までの製品紹介記事 3510 記事を使用した．91，92，94，95 年の 2627 記事を分析用データ，93 年の 874 記事を評価用データとし，製品単位，記事単位，項目単位の評価を行った．

| | |
|---|---|
| 抽出成功，対応成功 | 218 製品 |
| 抽出成功，対応失敗 | 8 製品 |
| 抽出超過，対応成功 | 20 製品 |
| 抽出超過，対応失敗 | 10 製品 |
| 抽出不足，対応成功 | 184 製品 |
| 抽出不足，対応失敗 | 120 製品 |
| 複数製品記事数 | 242 記事 |
| 製品数 | 560 製品 |

実験結果 1：製品単位の評価

| | 単数製品の記事 | 複数製品の記事 |
|---|---|---|
| 全記事数 | 631 記事 | 242 記事 |
| 製品数正解 | 598 記事 | 133 記事 |
| 製品数超過 | 33 記事 | 6 記事 |
| 製品数不足 | 0 記事 | 103 記事 |
| 情報抽出率 | 0.85 | 0.70 |

実験結果 2：記事単位の評価

| 抽出項目 | 正解情報数 | 再現率 | 適合率 |
|---|---|---|---|
| 販売元 | 884 | 0.93 | 0.95 |
| 発売日 | 678 | 0.91 | 0.93 |
| 製品種別 | 814 | 0.76 | 0.76 |
| 製品名 | 921 | 0.47 | 0.60 |
| 細分類 | 496 | 0.72 | 0.79 |
| 価格 | 1010 | 0.91 | 0.94 |

実験結果 3：項目単位の評価

実験結果 1 のから，抽出に成功した記事は対応付けも正しく行うことができた．抽出に失敗したものは，1 文が長い文章は KNP から正しい結果を得ることが困難であることが主な原因であった．また，固定パターンも価格も存在しない文章を抽出処理の対象としていなかったため，その文章にしか現れない一部の情報を漏らしたことも考えられる．

実験結果 2 では，単数の場合は高い割合で製品数を認識し，情報抽出率も高くなっている．しかし，複数の場合は全ての製品を抽出できない記事が多数あった．この原因は，実験結果 1 の抽出失敗とほぼ同様であった．

6 つの抽出項目の信頼性を測るために実験結果 3 の評価を行った．販売元，発売日のほとんどは固定パターンに係り，抽出が容易であった．また，価格は表記パターンでの抽出であるため，「○○円高」のような他との比較の価格を誤って選択した場合以外は正解となる．その他の項目は，抽出処理対象外の文に現れる可能性があることと，項目の定義が曖昧であること，文章中の位置が比較的自由に表現できるため，抽出に成功しても項目の割当に失敗することなどが抽出精度低下をもたらしていた．

## 5 まとめ

本論文では，これまでテンプレートを用いて別々に研究を行なってきた 1 記事中に 1 つの製品を紹介する記事からの抽出と，複数の製品を紹介する記事からの抽出を，構文解析結果を利用して統合処理を行なう手法を提案した．

テンプレート手法の統合による，(1) マッチするテンプレートが増えるため，多数の解候補があり抽出精度が低下する (2) 大量テンプレートとのマッチング，優先順位付けの処理によって，高速性が損なわれる，という問題点を，構文解析を行なう文章を限定し，必要最低限の解析結果のみを用いること，解がほぼ一意に決まるので優先順位付けが必要無いことで高速性を維持したうえで，解析結果を利用した高精度の抽出を行えるよう改良できた．

今後，抽出精度の向上を目指して，現在抽出対象外の文章に出現する抽出項目の分析を行い，より多くの情報を獲得すること，また，1 文中係り受けと階層関係だけでなく，記事中の出現位置とキーワードを分析し，1 記事単位のテンプレートを作成することで精度の低い項目の問題となっていた項目割り当てに利用することを考えている．

## 【参考文献】

[1] 井出裕二，藤吉誠，永井秀利，中村貞吾，野村浩郷：構造化テンプレートを用いた新聞記事からの製品情報抽出，情報処理学会研究報告 1997-NL-118

[2] 赤松順子，高尾宜之，永井秀利，中村貞吾，野村浩郷：複数製品の紹介記事からの製品情報抽出，情報処理学会研究報告 2000-NL-140

[3] 高尾宜之，永井秀利，中村貞吾，野村浩郷：複数製品の紹介記事からの製品情報抽出 -製品記述パターンの分析-，情報処理学会研究報告 1999-NL-129

[4] 黒橋禎夫，日本語構文解析システム KNP 使用説明書 version 2.0 b6，京都大学大学院工学研究科