自然言語処理 115－12
情報学基礎 43－4
（1996. 9. 12）

テンプレートを用いた新聞記事からの製品情報抽出システム

井出 裕二　藤吉 誠　永井 秀利　中村 貞吾　野村 浩郷

九州工業大学 情報工学部 知能情報工学科

E-mail: {ide,fujiyosi,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

本論文では，新聞における製品紹介記事からの情報抽出システムについて述べる．製品紹介記事は定型性が強く，それゆえ，表層的処理により記事中から情報を抽出することは十分に可能であると考えられる．本論文ではこの定型性をテンプレートとして表現し，このテンプレートとのパターンマッチングを基本とした情報抽出技法を提案する．このようなテンプレートを人手で作成するのは非常に多くの労力を要する．そこで，抽出されるべき情報を正解データとして与えた学習データ (記事) により，テンプレートを半自動的に作成する方法を示す．さらに，本手法に基づき作成された情報抽出システムにおいて，製品紹介記事2000 記事に対して行った製品情報抽出実験により，本手法の有効性を検証する．

Template-based Products Information Extraction from Newspaper Articles

Yuji Ide　Makoto Fujiyoshi　Hidetoshi Nagai
Teigo Nakamura　Hirosato Nomura

Department of Artificial Intelligence, Kyushu Institute of Technology

E-mail: {ide,fujiyosi,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

We propose a method for extracting products information from newspaper articles. Articles for products introduction have a fixed form. Therefore, we think that it is heighly probable to extract the information without deep semantic analysis. We represent characteristics of such a form as 'templates'. Our method is based on pattern matching with the templates. If templates have to be generated by hand, it will need a lot of manpower. So we propose the semi-automatic method for generating templates from learning data: articles and information which should be extracted. We describe the information extraction system with our method. And as a result of the experiments in information extraction from 2,000 articles on our system, we show the effectiveness of our method.

1 はじめに

ネットワークを介してのニュース記事の配送や，CD-ROM などの大容量記憶媒体を介しての情報の提供が一般的になったことにより，計算機可読な文書は爆発的に増加している．文書情報が大量にあればその中には有用な情報が多く含まれている可能性は大きいが，同時に自分にとって必要な情報を検索・抽出することは困難になる。したがって，大量の計算機可読な文書から目的とする情報を自動的に抽出する情報抽出システムの開発が望まれる．文章が定型性を持ちかつ抽出対象が明確である場合には、詳細な構文解析や深い意味解析を行わなくとも表層処理による簡易な処理で情報を高速に抽出することができると考えられる．新聞における製品紹介の記事は，定型的な文章で書かれていることが多いため，表層処理による実用的な情報抽出システムの実現が期待できる．

現在，我々は，新聞の製品紹介記事を題材とし，記事の定型性を利用することによって浅い言語処理で製品情報を高速に抽出することを目指している．具体的には，予め作成しておいたテンプレートと入力テキストを比較することにより，抽出されるべき情報を表す文字列を抽出する．よって，良質なテンプレートをいかに作成するかが重要な問題となるが，これを人手で作成するには非常に多くの労力を要し，またテンプレート作成者の主観が入ることにより一貫性が保てない恐れがあるため，作成の自動化を考える必要がある．

本論文では，学習データとして記事と，そこから抽出されるべき情報とを与え，人の支援の下に半自動的にテンプレートを作成する方法を提案する．またこの方法に基づき作成したテンプレートを用いて，情報抽出実験システムを構築する．最後に，このシステムを用いて抽出実験を行ない，その結果から字面処理による情報抽出の有効性を検討する．

2 テンプレート作成

2.1 テンプレート

抽出項目とパターンとの位置関係を記したものをテンプレートと定義し，テンプレートを定義するために，次の用語を用いる．

抽出項目： 抽出を試みる情報の内容
例 { 販売元 }，{ 製品種別 } など

固定パターン： 抽出情報の周辺に現れる特徴的な文字列
例 『売り出す．』，『販売した．』 など

パターン： 固定パターン，および抽出項目前後の語

ワイルドカード： 0個以上の文字からなる任意の文字列

L を抽出項目，P をパターン，W をワイルドカード，および C_i を $P_0W_1P_1W_2\cdots P_{n-1}W_nP_n$ とすると，テンプレート T は

$$T = C_0L_1C_1L_2\cdots C_{n-1}L_nC_n$$

と表す．なお，C_0 は先頭のパターン，C_n は後尾のパターンを省略できるものとする．テンプレートは1文単位で作成し，C と L が必ず交互に現われるものとする．また，テンプレートは記事の見出しと本文では文体がかなり異なるので，それぞれ別に作成する．

2.2 固定パターン

ここでは，文章中の定型的な表現である固定パターンを文章中から取り出す手順について述べる．

ここで，抽出項目に対応する，記事中から抽出されるべき情報を正解データと定義する．

1. 入力データから固定パターンの候補となる文字数が2から15までの文字列とその出現頻度を得る．
2. 1.で得た固定パターンの候補とそれらに対する直前・直後の隣接文字列集合から，各候補文字列のエントロピーを計算する．
3. 2.で得られたエントロピーの結果を高い方からソートする．
4. 3.の結果より，見出し，本文それぞれに対し固定パターンを取り出す．

各処理を具体的に説明する．

1.の処理では，入力データ中に2回以上出現した文字を列候補文字列として抽出する．この処理は長尾らの提唱した n-gram 統計による文字列抽出方法 [1] を利用する．

2.の処理では，隣接文字の直前の文字におけるエントロピーを前方エントロピー，直後の文字におけるエントロピーを後方エントロピーと

定義する．

この処理は下畑らの提唱したエントロピー基準 [2] を利用する．この方法によれば文字数や出現回数に依存することなく，文字列を比較できる．具体的な求め方としては，文字列 S の隣接文字集合を $W(S) = w_i|w_1 \cdots w_n$，隣接文字 w_i の生起確率を $P(S, w_i)$ とする時，エントロピー $H(S)$ は，以下の式で求められる．

$$H(S) = -\sum_{i=1}^{n} P(S, w_i) \cdot \log P(S, w_i)$$

例として“輸入販売”という文字列のエントロピーを求める．

出現頻度は２８回で，隣接文字情報は表１のようになる．この結果を上記のエントロピーを求める式に当てはめると，前方エントロピーは 1.413，後方エントロピーは 1.394 となる．

表１ “輸入販売”の隣接文字情報

前接文字	出現回数	後接文字	出現回数
’ ’	3	会	2
で	1	す	9
を	3	な	1
の	16	を	12
—	1	に	3
に	2	事	1
ら	2		

3．の処理では，前方エントロピーと後方エントロピーを比較して，低い方を有効値とする．

4．の処理では，以下の条件に当てはまるものを除去する．

- 語頭にならない文字（「ん，ゃ，っ」など）から始まる文字列
- エントロピーがどちらか１つでも０になる文字列

また，テンプレートの抽出項目の抽出においては文末表現が抽出精度に大きく寄与する．よって「本文」においては今回は文末表現のみを固定パターンとし以下のものと定める．また「見出し」は文として成り立っていないものが多くあるため，固定パターンを文末表現のみに限定しないほうが抽出精度が上がる．よって「見出し」においては固定パターンを以下のものと定める．

見出し 助詞で始まっている文字列は，助詞を取り除く．また断片的文字列は取り除く．

本文 文末表現のみ．助詞で始まっている文字列は，助詞を取り除く．また断片的文字列は取り除く．

2.3 テンプレート作成

学習用記事からテンプレートを作成する手順を述べる．

抽出項目に対応する情報を表す文字列を抽出情報と呼ぶ．また，２．２節で求めた固定パターンをいくつ用いるかは，人手によって判断する．

1. 人手によって学習用記事から抽出情報を抜きだし，正解データを作成する
2. 学習用記事を形態素へ分割する
3. 記事中の抽出情報を抽出項目に置き換える
4. 抽出情報と固定パターン以外の単語をワイルドカードに置換する．ただし，抽出項目の直前直後の単語は置換せずに残す
5. 連続したワイルドカードを一つにまとめる

この方法によるテンプレートの作成例を次に示す．文１に対して正解データ１を与えることにより，テンプレート１が生成される．固定パターンである「売り出す．」はワイルドカードに置換されない．テンプレート１において，＊はワイルドカード，{} で囲まれた部分は抽出項目名を表す．

文１ 日立製作所は家庭用ビデオＣＤプレーヤーを来年一月をメドに売り出す．

正解データ１ 販売元: 日立製作所，製品種別: ビデオＣＤプレーヤー，発売日: 来年一月

固定パターン 売り出す．

テンプレート１ {販売元}は＊用{製品種別}を{発売日}を＊に売り出す．

3 情報抽出システム

今研究で作成した製品情報抽出システムについて述べる．なお，本システムにおいて，抽出項目は製品種別，製品名，販売元，価格，発売日の５つとする．

3.1 テンプレートを用いた情報抽出

テンプレートを用いた情報抽出処理について述べる．

パターンは抽出項目またはワイルドカードの前後に現れ，パターンとパターンに挟まれた文字列は，対応するのがワイルドカードではなく抽出項目ならば，抽出情報として抽出される．テンプレートに記述されている順番ですべてのパターンと入力文とのパターンマッチに成功すると，パターンには挟まれた抽出項目の抽出情報を得ることができる．

図 1の例では，テンプレートの中のパターン『は』『の』『「』『」を』『に発売する．』が入力文中に存在し，かつ出現順序もテンプレートと同じであるため，４つの抽出項目についての具体的な抽出情報を得ることができた．

パターンマッチは，記事を句点で分割して，１文ごとに行うが，何文目かによって抽出情報を含む確率が異なり，本文部の１文目，２文目，最終文，および見出しに抽出情報が含まれている確率が高い．よって，本実験システムにおいてはパターンマッチを行う文の順番は

- 「本文」の１文目 → 「本文」の２文目 → 「本文」の３文目 → 見出し → 「本文」の残りの文を上から の順でパターンマッチ

となる．なお，記事の本文部は最低２文，最高１６文からなる．

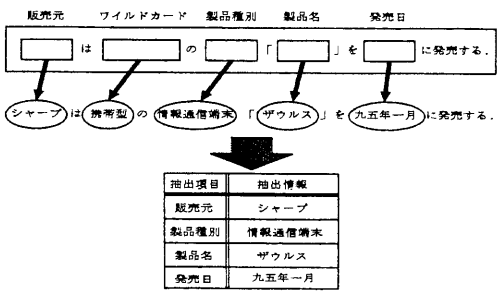

図 1: テンプレートを用いた情報抽出

3.2 入力テキストの形式

入力テキストは，１行目 (見出し) に記事の概要，２行目以降 (本文) が記事の詳細という形式をとる．入力テキストの例は次のようなものである．

ダイキン工業，家庭用空気清浄器，白とグレーの２機種．

ダイキン工業は一般家庭用の空気清浄器「クリエール」をモデルチェンジし，外観が白とグレーの二機種を来年二月に発売する．ちりを電気的に捕集する清浄器で，「花粉モード」などちりの質に応じた運転方式を三種類設定した．機種名は「ＡＣＥＦ２５Ｄ」で，高さ三十七センチ，幅四十七センチ，奥行き十三センチ．価格は四万四千八百円で，年間三万五千台以上の販売を目指す．

3.3 抽出項目の特徴

抽出項目を示す情報のなかで「価格」と「発売日」を示す情報は少ないパターンで表現できる．以下にそれぞれの特徴について説明する．

- 価格：価格を示す情報は主に「100 円」などのように数字が続き最後の文字が通貨単位で終わる．このように価格は少ないパターンで表現できる．

- 発売日：発売日を示す情報は主に「2 月 16 日」などのように数字で始まり最後の文字が”日”などで終わることが多い．このように発売日は少ないパターンで表現できる．

本システムにおいては「価格」と「発売日」のパターンを日本経済新聞 1994 年度版製品紹介記事 2,000 記事から作成した．

3.4 情報抽出アルゴリズム

見出し用テンプレート集合が HEAD-TEM[1] ～ HEAD-TEM[m]，本文用テンプレート集合が BODY-TEM[1] ～ BODY-TEM[n] までであるものとする．抽出アルゴリズムの概要を図 2に示す．

1. 記事を句点で分割する

2. 各抽出項目に対応する抽出情報が得られるまで，決められた順番で文に以下の処理を行う

   Step1 テンプレートを用いたパターンマッチ
   文が「見出し」ならば HEAD-TEM[1] ～ HEAD-TEM[m] の順番で文にパターンマッチさせ，
   文が「本文」ならば BODY-TEM[1] ～ BODY-TEM[n] の順番で文にパターンマッチさせる． Step2 へ．

Step2 抽出情報のチェック
テンプレートによって抽出された文字列抽出項目ごとに以下のように処理する

(a) 製品種別：あらかじめ作成した「製品種別」データを参照し，「製品種別」として抽出された文字列がその中に含まれていれば抽出情報と決定する

(b) 製品名：抽出された文字列をそれぞれの抽出項目に対する抽出情報とし決定する

(c) 販売元：あらかじめ作成した「販売元」データを参照し，「販売元」として抽出された文字列がその中に含まれていれば抽出情報と決定する

(d) 価格：「価格」として抽出された文字列が価格のパターンにマッチすれば抽出情報と決定する

(e) 発売日：「発売日」として抽出された文字列が発売日のパターンにマッチすれば抽出情報と決定する

3. 抽出情報が一つも得られない場合は，各抽出項目のチェックなしに，最初にテンプレートから抽出された文字列を抽出情報と決定する

4. 各抽出項目ごとに最初に決定された抽出情報を出力する

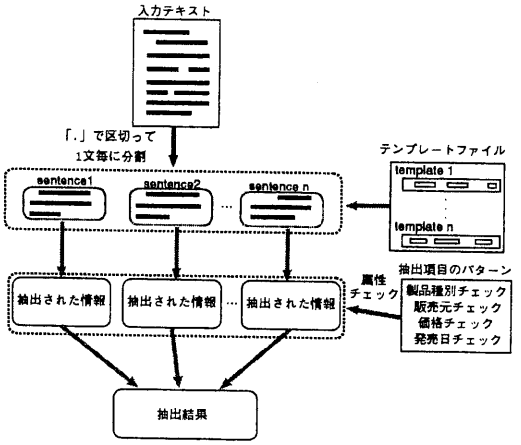

図 2: 情報抽出アルゴリズム概要

4 情報抽出実験とその評価

本章では，前章で作成した実験システムを用いて新聞中の製品情報欄から表層処理により製品情報を抽出する実験を行い，その結果を考察する．

4.1 実験題材

実験題材としては，日本経済新聞 1994 年度版から製品紹介記事を２，０００記事を用いた．

4.2 実験の手順

今実験においては実験１として，2,000 記事を 200 記事づつＡ～Ｊの 10 グループに分け，１グループを検証データとし，その他の 9 グループを学習データとし情報抽出を行った．また併せて実験２として，2,000 記事すべてを学習データとし，作成したテンプレートを用いて同記事からの情報抽出を行った． まず実験は，１．テンプレートの作成，２．作成したテンプレートを用いた情報抽出実験についてそれぞれ行う．文からテンプレートを作成するためには，その文に含まれている抽出情報が分かっていなければならない．したがって，あらかじめ各記事に対する抽出情報の正解データを用意しておく．

テンプレート作成の手順

各グループに対して以下の処理を行う．

1. 学習データから各記事の正解データを含む文を抜き出しテンプレート作成データとする

2. テンプレート作成データからそれぞれ見出し用，本文用のテンプレートを作成する

3. 作成したテンプレートの中で同一のものは一つにまとめる

4. 実験システムを用いて各テンプレートでパターンマッチングを行いテンプレート作成データから情報抽出を行う

5. テンプレートとマッチした文の数と，正解データを用いてそのテンプレートから正しい抽出情報が得られた文の数を求める

6. テンプレートとマッチした文の数に対する正しい抽出情報が得られた文の数の割合が高い順にテンプレートを並べ換える

情報抽出実験の手順

Ａ～Ｈの各グループの検証データに対して，対応したテンプレート作成データで作成したテンプレートを用いて以下の処理を行う．

1. 実験システムを用いて検証データ各記事から情報抽出を行う

2. 抽出した抽出項目の内容のうち，「価格」，「発売日」については価格パターン，発売日パターンにマッチした場合のみ抽出情報として扱う．また「製品種別」，「販売元」については，あらかじめ作成した製品種別，販売元データベースに登録されている場合のみ抽出情報として扱う．

3. 正解データを用いて抽出結果について考察する

4.3 実験の結果

本実験の結果として1．作成されたテンプレートの数，2．情報抽出結果 を以下に示す．

1．作成されたテンプレートの数

実験1の結果として，各グループの学習データから作成されたテンプレートの数の平均は次のようになった．

見出し 1174.4 個

本文 1521.6 個

また，実験2として，2,000 記事すべてを学習データとした場合は，テンプレートの数は次のようになった．

見出し 1296 個

本文 1672 個

2．情報抽出結果

表2に抽出結果を示す．ここで抽出率は以下の式により計算される．

また，正しく抽出されなかった情報でも，その情報に正しい抽出情報が含まれている場合，すなわち余分な文字列を含んで抽出された場合は，今後の処理(フィルターをかけるなど)で正しい抽出情報を得られる可能性がある．よって，正しい情報に余分な文字列を含んで抽出した情報も正解とみなし，その抽出した割合を準抽出率として以下の式で定義する．

抽出率 ＝ 正しく抽出された記事の数 / 抽出項目の情報が記述されている記事の数

準抽出率 ＝ (正しく抽出された記事の数 ＋ 正しい情報に余分な文字列が付随して抽出された記事の数) / 正解が分かっている記事の数

表2 抽出結果 (%) A: 抽出率 B: 準抽出率

	実験1		実験2	
抽出項目	A	B	A	B
製品種別	66.18	79.88	91.06	93.31
製品名	64.07	65.19	65.53	66.41
販売元	89.78	92.86	95.82	96.78
価格	66.93	69.61	79.36	80.90
発売日	80.06	87.27	87.69	91.62

4.4 実験の考察

今実験の考察として1．テンプレートの作成，2．情報抽出結果について述べる．

1．テンプレート作成

作成したテンプレートの数を「見出し用」と「本文用」で比較してみると，テンプレート候補文からテンプレートは以下の数だけ作成される．

見出し用 1981 文 → 1296 個

本文用 4963 文 → 1673 個

上記の結果から，「見出し用」よりも「本文用」のテンプレートの方が一般化の度合が大きいことが分かる．これは，以下の理由による．

- 「見出し」は，特殊な文体であるものが多いため
- 「見出し」は，テンプレートを作成するとき，固定パターンとして文末表現だけでなく，文中の定型的な表現を用いているため

また，今回提案したテンプレート作成方法では，以下のように文1からテンプレート1が作成される．

文1 平和食品工業は七月二十一日に全国で一斉に販売を始める．

テンプレート1 {販売元1} は {発売日1}に ＊ 販売を始める．

このテンプレート1を文1に適用すると，本当は「発売日」は『七月二十一日』を抽出したいのだが，『七月二十一日に全国で一斉』が抽出されてしまう．これは抽出項目をパターンとパターンの間を最長一致で抽出するためである．

この解決法としては，

- 抽出項目の後ろのパターンを増やす
  テンプレート1 {販売元 1} は {発売日 1}に全国 *販売を始める.
- ワイルドカード項を増やす
  テンプレート1 {販売元 1} は {発売日 1}に* に* 販売を始める.

が挙げられる. これらの処理は今後の課題である.

2. 情報抽出実験結果

- 情報の抽出に失敗した原因
  * パターンマッチを行って最初に適用が成功したテンプレートの抽出した情報を即抽出情報として採用しているため, より適したテンプレートが後から適用されてもその情報は正しい抽出情報として採用されない. 特に「製品名」においては, 2番目以降に正しく抽出された記事の数は319記事あった. 抽出情報の決定法の再考が必要である.
  * テンプレート作成の考察でも触れたが, テンプレートは用意されているが, 望んでいたところにマッチしない.
- 5つの抽出項目の中では「製品種別」, 「製品名」および「価格」が検証データからの抽出率が低い. そのうち, 「製品種別」と「価格」の実験1の抽出率は実験2の抽出率よりも10%以上低い. これは, 検証データの「製品種別」と「価格」を含む文にマッチするテンプレートがないためで, 今後, 学習データを増やすことにより解決できる. 「製品名」は検証データ, 学習データからの抽出率ともに低い. これは, 「製品名」の場合, 情報抽出する際に何も制約をつけずに, テンプレートで抽出された文字列をそのまま抽出情報としているためである. 今後の課題として何か対処法を考える必要がある.
- テンプレートからのパターンマッチによって抽出情報として得たい情報の他に周辺の文字列も含めて抽出された場合, 今後の研究によってそれらの文字列から正しい抽出情報が得られるようにすることは可能である. よって今回の実験の成果として準抽出率まで見てみると, 実験1の結果において各項目の平均の抽出率は78.96%であり, 特に「販売元」の結果は92.86%と高い結果を示した. よって, 今実験で用いた実験システムによって, 定型性を有する文章に対する浅い処理による情報抽出は有効であると言える.

4.5 テンプレートの縮約及び実験

情報抽出処理においては, テンプレートの数が処理時間に大きく関わっており, 数が少ないほど高速に抽出処理が行える. そこで学習用記事から作成されたテンプレートに対し, 統合可能と思われるものをまとめることで抽出率を落さずにテンプレートの数を絞り, 処理時間の短縮を目指す.

4.5.1 テンプレート縮約方法

本研究では, テンプレート同士の包含関係を調べて, テンプレート数の削減を行う (図3). 具体的には, 作成したテンプレートを用いて各学習記事から情報抽出処理を行い, どのテンプレートがどの文において正しい情報抽出ができたかを記録する. その結果, あるテンプレートAを適用可能であった文集合が, 別のテンプレートBを適用可能であった文集合に包含されているならば, テンプレートBはテンプレートAをより一般化したものであるとみなし, テンプレートAを削減する.

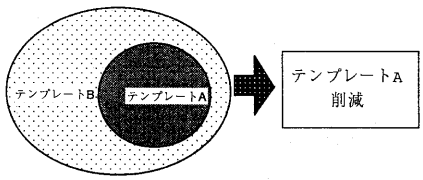

図 3: 包含関係によるテンプレートの削減

4.6 縮約したテンプレートを用いた情報抽出実験

ここでは4.5節で述べた方法で縮約したテンプレートを用いた情報抽出実験について述べる.

4.6.1 実験方法

縮約したテンプレートを用いて，4．2節で述べた方法で情報抽出実験を行う．

4.6.2 実験結果

テンプレートを縮約した結果，各グループのテンプレートの平均数は以下のようになった．

	縮約前		縮約後
見出し	1174.4 個	→	866.6 個
本文	1521.6 個	→	1203.4 個

また，表3に縮約前と後の抽出結果を示す．

表3 抽出結果 (%)

	縮約前		縮約後	
抽出項目	A	B	A	B
製品種別	66.18	79.88	66.23	79.92
製品名	64.07	65.19	61.62	62.56
販売元	89.78	92.86	90.69	93.66
価格	66.93	69.61	66.70	67.55
発売日	80.06	87.27	80.06	87.01

A: 抽出率 B: 準抽出率

4.6.3 実験の考察

今実験の情報抽出処理においては，処理時間はテンプレートの数が増えればそれに伴って増加する．今回のテンプレート削減処理によってテンプレートの数は，「見出し」，「本文」あわせて平均約2,700個から平均約2,000個に削減され処理時間は短縮された．

また，テンプレート削減前の実験結果とテンプレート削減後の実験結果を比べてみると結果はほとんど変わりはなく，項目によっては抽出率が上昇したものもある．

以上の事より，今回使用したテンプレート削減アルゴリズムは有効であったといえる．

5 おわりに

定型的な文章で書かれていることが多い新聞記事中の製品紹介記事を対象とした情報抽出システムの構築について提案した．

また，情報抽出システムで用いるテンプレートの作成について述べた．学習用記事と正解データおよび，固定パターンから，テンプレートを自動的に作成することができ，またテンプレートの包含関係を考慮することにより，抽出率を下げずにテンプレートを縮約することができた．

最後に，日本経済新聞1994年度版中の製品紹介記事2,000記事を用いて，情報抽出の実験を行い，ここで述べた情報抽出方式の有効性を検証した．　今後の課題としては，以下のものが挙げられる．

- 処理時間短縮のため，テンプレートの一般化
- 抽出率向上のため，抽出システムの改良
- 『製品種別』にあわせて，抽出項目を決定する

謝辞

本論文で使用したテキストデータは，「日本経済新聞記事データ CD-ROM(1994版)」を使用した．使用を許可して下さった日本経済新聞社，および日経総合販売(株)に深く感謝致します．

参考文献

[1] 長尾眞，森信介："大規模日本語テキストのnグラム統計の作り方と語句の自動抽出"，情報処理学会自然言語処理研究会報告，96-1，pp.1–8，(1993).

[2] 下畑さより，杉尾俊之，永田淳次："隣接文字の分散値を用いた定型表現の自動抽出"，情報処理学会自然言語処理研究会報告，110-11，pp.71–78，(1995).

[3] 松尾比呂志，木本晴夫："抽出パターンの階層的照合に基づく日本語テキストからの内容抽出法"，情報処理学会論文誌，Vol.36，No.8，pp.1838–1844，(1995).

[4] 井出 裕二："テンプレートを用いた新聞記事からの製品情報抽出システム"，九州工業大学卒業論文（1996）．

[5] 藤吉　誠：テンプレートを用いた新聞記事からの製品情報の抽出，九州工業大学卒業論文（1995）．