自然言語処理 122－10
（1997. 11. 20）

単一項目テンプレートによる新聞記事からの製品情報抽出

井出 裕二　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,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

我々は，抽出すべき情報とその周辺の文字列との関係を記した“テンプレート”を用いて，字面処理による情報抽出処理の研究を行っており，新聞の製品紹介記事を題材とした実験で，テンプレートを用いた抽出処理の有効性を確認してきた．しかし，従来のテンプレートの形式では，１つのテンプレートに複数の抽出項目を含んでいたため，抽出率(再現率)を上げるためには，多数のテンプレートを用いなければならず，結果として処理時間が増加するという問題があった．

本論文では，この問題を解決する方法として，テンプレート中に記述する抽出項目を単一の項目のみに制限した単一項目テンプレートを用いた情報抽出処理について述べる．そして，実際の製品紹介記事を用いて行なった抽出実験結果と従来のものを比較することによって，単一項目テンプレートによって処理速度，再現率を向上させることができることを示す．

Information Extraction with Single Item Template from Newspaper Articles

Yuji Ide　Hidetoshi Nagai
Teigo Nakamura　Hirosato Nomura

Department of Artificial Intelligence, Kyushu Institute of Technology

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

We have been researching a textual analysis method for information extraction from newspaper articles with templates which describe the relationship between information to be extracted and its surrounding strings and have confirmed the effectiveness of our method by experiments.However, as each template contains more than one item to be extracted, the set of templates becomes large and it takes a long time to match an input text with all templates.

In this paper, we propose “single item templates” that contain only one item to be extracted to reduce the number of templates and processing time. We show the experimental result to illustrate the effectiveness of our method with single item templates.

1 はじめに

ネットワークの普及により，膨大かつ多種多様な情報が流通するようになった現在，それらの情報の中から必要な情報だけを人間の手作業で取り出してくることはもはや不可能になりつつある．このような状況下において，計算機可読な文書情報を計算機で自動的に処理し，必要とする情報のみを取り出してくることができれば，氾濫する情報も効率的に管理することができ，情報の価値を落とすことなく適切な利用が可能となる．このように，計算機による情報抽出システムが実現することによって得られるメリットは非常に大きいため，現在盛んに研究されている [4][5][6][7].

筆者らは，大量の計算機可読文書から目的とする情報を高速に抽出する研究を行なっており，これまでに，新聞の新製品紹介記事を対象として，記事の記述形式の定型性を利用した製品情報抽出方式の有効性について報告を行なってきた [1][2][3]．そこで用いている基本的な手法は，抽出すべき項目とその周辺の文字列との関係を記述したテンプレートと入力記事とのマッチングによるものであるが，従来のテンプレートの形式では，一つのテンプレートに複数の抽出項目を含んでいたため，抽出率 (再現率) を上げるためには，多数のテンプレートを用いなければならず，結果として処理時間が増加するという問題があった．この問題を改善する一つの手法は，テンプレートを構造化して無駄なマッチングを削減する手法であり，これは文献 [2] において報告した．

本稿では，この問題を解決する別法として，テンプレート中に記述する抽出項目を単一の項目のみに制限した単一項目テンプレートを用いた情報抽出処理について述べる．そして，実際の製品紹介記事を用いて行なった抽出実験結果と従来のものとを比較することによって，少数のテンプレートによって再現率を向上させることができることを示す．

2 テンプレートを用いた情報抽出

2.1 テンプレート

以下では，情報抽出処理のためのテンプレートを定義する．まず，テンプレートの定義のための用語を定める．

抽出項目: 抽出を試みる情報の内容を表すラベル (「製品種別」，「製品名」，「販売元」，「価格」，「発売日」の 5 項目)

パターンマッチングにおいてはワイルドカードと同様に機能する

抽出情報: 抽出項目に対応する情報を表す文字列

パターン: パターンマッチングの対象となる文字列長 1 文字以上の文字列

固定パターン 抽出対象に頻出する特徴的な文字列 (“発売する”，“販売する” など)

ワイルドカード: パターンマッチング上，文字列長 0 以上の任意の文字列とマッチしうるシンボル

テンプレートは次のように定義する．

L を抽出項目，P をパターン，W をワイルドカードとしたとき，テンプレート T を

$$T = C_0L_1C_1L_2\cdots C_{n-1}L_nC_n$$
$$(C_i = P_0W_1P_1W_2\cdots P_{m-1}W_mP_m)$$

と定める．なお，文頭の C_0 および文末の C_n は空文字列であってもよい．テンプレートは 1 文単位で作成し，C と L は必ず交互に現れるものとする．

ワイルドカードを導入することにより，テンプレートをより一般化することができる．テンプレートの一般化の長所は，テンプレート集合の削減につながり，抽出処理に要する時間を短縮することができる点である．

以後，テンプレートを表現する場合，抽出項目は {item} の形式で表す．item は抽出項目名である．ワイルドカードは * で表す．

2.2 テンプレート作成

テンプレート作成用記事からテンプレートを作成する手順は以下の様になる．

1. 人手によってテンプレート作成用データからテンプレートに変換する文を選択する
2. 記事中の抽出情報を抽出項目に置き換える
3. 抽出情報前後の 1 文字と固定パターン以外の文字列をワイルドカードに置換する

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．

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＊売り出す.

2.3 1文からの情報抽出

入力文がテンプレートとのパターンマッチングに成功すると，パターンに挟まれた入力文中の文字列が対応する抽出項目の抽出情報として抽出することができる.

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.

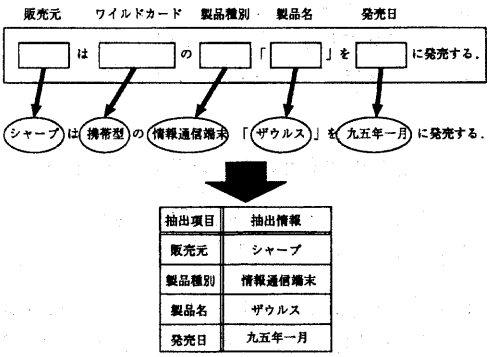

図 1: 1文からの情報抽出

2.4 1記事からの情報抽出

テンプレートを用いた1記事からの情報抽出処理は，以下の様な手順で行われる.

1. 記事を句点で分割する.
2. 各文に全てのテンプレートをマッチングさせ，マッチした場合は以下の処理をを行う.

(a) 抽出項目に対する制約を用いて，テンプレートとマッチングで得られた文字列をチェックする [2].

(b) 1つのテンプレートから得られた全ての文字列が制約チェックを通ったならば，それぞれの抽出項目の解候補に加える.

3. 各抽出項目の解候補に対して優先順位付けを行い，それぞれ抽出情報を決定する.

図2に処理の全体の流れを示す. 従来，優先順位付けは各候補文字列に対して，その文字列を抽出したテンプレートの個数をスコアとし，そのスコアが最も高いものを抽出情報と決定してきた.

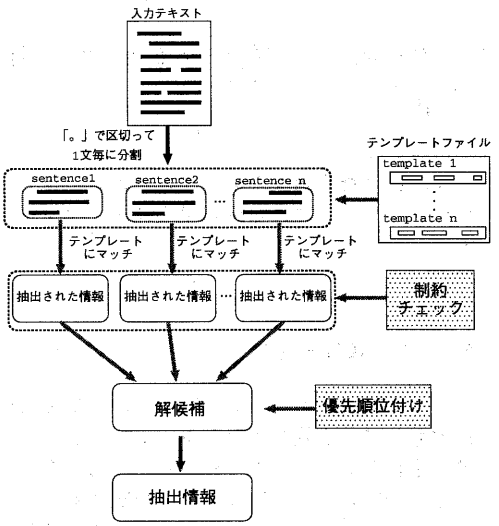

図 2: 1記事からの情報抽出

2.5 従来の研究の問題点

図1のように，一つのテンプレートに複数の抽出項目を含むテンプレートを以後，複数項目テンプレートと呼ぶ. これまでの実験では，テンプレート作成用サンプルとほぼ同数の複数項目テンプレートが作成され，抽出処理にかかる時間が大きくなり，しかもそれでもなお，テンプレート不足のため正しい情報抽出ができない記事があるという状況にあった.

この原因は，複数項目テンプレートでは，抽出項目，固定パターン，抽出情報の前後の1文字の組合せ (表1) を網羅してテンプレートを作成してしまうためである.

そこで，抽出精度を下げずに，テンプレートの一般性を高め，かつ，テンプレート数を削減することが必要となる.

	前1文字	後1文字
製品種別	180個	113個
製品名	17個	21個
販売元	41個	41個
価格	83個	23個
発売日	39個	28個

表 1: 抽出情報の前後に表れる1文字の異なり数

3 単一項目テンプレートを用いた情報抽出

前に述べた問題点を解決するために，単一項目テンプレートによる情報抽出法を，以下に示す．

3.1 単一項目テンプレート

1つのテンプレートに1個以上の抽出項目が記述されている複数項目テンプレートに対し，単一項目テンプレートでは，抽出項目としては1個しか含まない．このように制限することによって，テンプレート記述における抽出項目の周辺の自由度が増し，組合せの数が減少する．そしてその結果として，全体のテンプレート数を減少させることができる．また，単一項目テンプレートは複数項目テンプレートよりも一般性が高いため，複数項目テンプレートに比べ，テンプレート不足のため正しい情報抽出ができないという状況も改善される．

その反面，単一項目テンプレートは一般性が高いため複数項目テンプレートよりもマッチングしやすく，その結果，解候補が増大する．そこで単一項目テンプレートを用いる情報抽出処理では，次章で述べる解の選択処理がより重要となってくる．

ここで，テンプレート作成用データが十分でない場合は，単一項目テンプレート数は複数項目テンプレート数よりおおくなってしまう．下の例の場合は，複数項目テンプレート2個に対して，単一項目テンプレートは4個作成され，テンプレート数は逆に増えてしまう．

複数項目テンプレート

[{販売元}は，{発売日}から発売する．
{販売元}が*，{発売日}に販売する．]

↓

単一項目テンプレート

[{販売元}は，*から発売する．
*，{発売日}から発売する．
{販売元}が*販売する．
*，{発売日}に販売する．]

(※) 下線部は固定パターン

しかし，テンプレート作成用データの個数が十分多かった場合，単一項目テンプレートは一般性が高いため，同じテンプレートが多く作られ，結果的にテンプレートの個数は複数項目テンプレートよりも少なくなる．

さらに次に，抽出精度をほとんど下げずにテンプレートを削減する方法として，テンプレート同士の包含関係を用いた削減法，および，パターンの統合による削減法を提案する．

3.2 包含関係を用いたテンプレートの削減

テンプレート同士の包含関係を用いて，テンプレートを削減する方法を示す．

1. 全てのテンプレートをテンプレート作成用データの全ての文にマッチさせ，マッチしてかつ情報抽出が成功した文を記憶しておく

2. 全てのテンプレート同士を比較して，あるテンプレートの情報抽出が成功した文集合が，別のテンプレートの情報抽出が成功した文集合に包含されているならば，そのテンプレートを削減する (図3)

上の例では，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．

この方法では，情報抽出が成功した文集合の包含関係を用いているため，テンプレートの削減によって，テンプレートにマッチしない文が増えることはない．

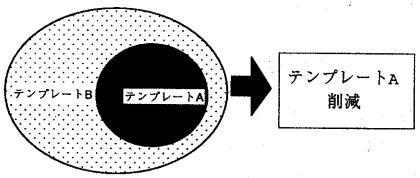

図 3: 包含関係によるテンプレートの削減

3.3 パターンの統合によるテンプレートの削減

テンプレート中のパターンを統合することによってテンプレートを削減する方法を以下に示す.

削減方法は, 以下の条件を満たすテンプレート同士の1つだけ異なるパターンを統合することによって行う.

- 抽出項目とワイルドカードの並びが同じ
- パターンが1つだけ異なる

$$\left[\begin{array}{l} *が\{製品種別\}は*発売する. \\ *で\{製品種別\}は*発売する. \\ 開発した\{製品種別\}は*. \\ 発売した\{製品種別\}は*. \\ \{製品種別\}は*. \end{array}\right]$$

$$\downarrow$$

$$\left[\begin{array}{l} *(が|で)\{製品種別\}は*発売する. \\ (開発した|発売した|)\{製品種別\}は*. \end{array}\right]$$

この方法においては, 統合したパターンを分解すればまたもとのテンプレートに戻るため, 削減する事によってテンプレートにマッチしない文が増えることはない.

4 解の選択

単一項目テンプレートは複数項目テンプレートに比べ一般性が高いため, 製品紹介記事からマッチングによって得られる各抽出項目の解候補の数は, 複数項目テンプレートを用いる場合よりも多くなる. そのため, 解候補に対し優先順位付けを行って抽出情報を選択する処理が一層重要となる.

従来, "すべてのテンプレートは同じ確率で情報を抽出する"と考えて, マッチしたテンプレートの個数によって解候補の優先順位付けを行ってきた. しかしテンプレートによっては, マッチする文は少ないが高い精度で正しく情報を抽出するものもあれば, マッチする文は多いけれども抽出精度は低いというものもある.

そこで今回, 我々は以下の方法で全てのテンプレートに重みを付けて, それを用いて解候補の中から抽出情報を決定した.

重みの決定法

1. テンプレート集合の各テンプレートを学習用データの全ての文にマッチさせ, マッチした文及び, マッチしてかつ情報抽出が成功した文の個数を記憶しておく.

2. 以下の式で重みを決定する.

$$重み = \frac{マッチしてかつ情報抽出が成功した文の数}{マッチした文の数}$$

4.1 解の選択法

各抽出項目の解候補の中の一つ一つの文字列は, 1個以上のテンプレートから抽出される. そこで, その一つ一つの文字列を抽出したテンプレートの重みの合計を, その文字列のスコアとする. そして, そのスコアが一番大きい文字列を, その記事の抽出項目の抽出情報と決定する.

例えば, テンプレート t_1,t_2,t_3,t_4 のそれぞれの重みがそれぞれ0.9, 0.7, 0.4, 0.3であるとき, 以下の表2の場合では, 文字列"A"が抽出情報となる.

文字列	マッチしたテンプレート	スコア
A	t_1,t_2	1.6
B	t_2,t_3,t_4	1.4

表 2: 抽出文字列に対するスコア付け例

5 実験と考察

今回提案した単一項目テンプレートと従来の複数項目テンプレートを用いて，新聞の製品紹介記事からの情報抽出実験を行い，その結果を考察する．

5.1 実験対象

実験では，日本経済新聞1994年版の中の製品紹介記事2000記事を用いた．1記事中に文は平均6.8文(13668文／2000記事)あり，また，文字数は平均303.2文字(606051文字／2000記事)，最大1131文字，最小69文字であった．1文当たりの文字数は平均44.3文字(606051文字／13668文)であった．

製品紹介記事2000記事のうち1800記事をテンプレート作成用データとして使用し，残り200記事を評価用データとして使用した．

5.2 実験の手順

単一項目テンプレート，複数項目テンプレート(従来)に対して，以下の実験を行う．

テンプレート作成 次の3種類のテンプレート集合を作成し，その数を比較する．

T1 テンプレート作成用記事1800記事から作成したテンプレート集合

T2 T1のテンプレート集合を包含関係を用いて削減したテンプレート集合

T3 T2のテンプレート集合のパターンを統合して削減したテンプレート集合

処理速度 Pentium200MHzマシン上で，1記事当りの情報抽出処理時間を計測する．

情報抽出精度 作成したテンプレートを用いて評価用データに対して情報抽出処理を行い，解候補の個数(曖昧性)と抽出精度を求める．抽出精度は以下に示す再現率と適合率を用いて行う．

$$\text{再現率} = \frac{\text{正しく抽出された記事の数}}{\text{抽出項目の情報が記述されている記事の数}}$$

$$\text{適合率} = \frac{\text{正しく抽出された記事の数}}{\text{情報を抽出した記事の数}}$$

5.3 実験結果

テンプレート作成

表3は，テンプレート作成用データ1800記事から作成された単一項目テンプレート，及び複数項目テンプレートの数を示す．

	単一項目	複数項目
T1	3544	2779
T2	1326	1997
T3	605	1412

表 3: テンプレートの作成結果(個)

処理速度

表4は評価用データに対して情報抽出実験を行ったときの，1記事の平均処理時間を示す．

	単一項目	複数項目
T1	5.31	5.16
T2	2.34	4.02
T3	1.25	3.29

表 4: 処理速度の測定結果(秒)

抽出精度

評価用データに対して情報抽出実験を行ったときの結果を表5,6,7,8に示す．表5は各抽出項目の解候補の平均数を示し，表6,7,8は抽出結果の再現率，適合率を示す．

5.4 考察

5.4.1 テンプレート作成

T1では単一項目テンプレートは複数項目テンプレートよりも多く作成されるが，削減処理を行うことによってT2では約2/3，T3では1/2

	単一項目	複数項目
T1	14.68	6.13
T2	14.26	6.03
T3	9.99	5.60

表 5: 解候補数の平均 (個)

	再現率		適合率	
抽出項目	単一	複数	単一	複数
製品種別	62.00	56.50	62.00	56.50
製品名	93.87	92.64	85.00	84.83
販売元	89.23	89.74	89.23	89.74
価格	90.24	89.63	95.48	94.84
発売日	95.78	86.14	89.83	88.27

表 6: テンプレート T1 を用いた抽出結果 (%)

	再現率		適合率	
抽出項目	単一	複数	単一	複数
製品種別	63.50	56.50	63.50	56.50
製品名	90.18	88.96	81.67	81.46
販売元	87.69	89.23	87.69	89.23
価格	88.41	87.80	93.55	92.90
発売日	93.37	86.14	87.57	89.38

表 7: テンプレート T2 を用いた抽出結果 (%)

	再現率		適合率	
抽出項目	単一	複数	単一	複数
製品種別	56.50	50.00	56.50	50.00
製品名	88.34	81.60	80.00	74.72
販売元	71.28	80.51	71.28	80.51
価格	87.80	88.41	93.51	94.16
発売日	91.57	83.13	86.36	86.25

表 8: テンプレート T3 を用いた抽出結果 (%)

以下になっている．これは単一項目テンプレートは複数項目テンプレートよりも抽出項目と周辺の固定パターンの自由度が高いために，多くのテンプレートを 1 つにまとめあげることができるためである．

また，T1 から T2 への包含関係によって削減されたテンプレートを調査してみるとほとんどが，包含されるテンプレートのパターンがワイルドカードに吸収される事によって削減されている．

5.4.2 処理速度

単一項目テンプレートと複数項目テンプレートでは，単一項目テンプレートの方が 1 文あたりのマッチング時間がかからないため，T1 ではテンプレート数は単一項目テンプレートの方が多いが，処理速度はほとんど変わらない．

単一項目テンプレートでテンプレート数と処理時間の関係では，T1 と T3 を比較してみるとテンプレート数は約 1/6 になっているが処理速度は約 1/5 にしかなっていない．これは，テンプレート削減処理によって一般性が高くなるため，1 個のテンプレートがマッチする文の数が増えてしまうためである．

5.4.3 抽出精度

抽出精度については以下の 3 つの立場から考察する．

単一項目テンプレートと複数項目テンプレート

抽出精度を比較すると，T1,T2,T3 のほとんどの項目で単一項目テンプレートのほうが

高い結果を示している．これは，単一項目テンプレートは複数項目テンプレートよりも一般性が高いため，複数項目テンプレートに比べ，テンプレート不足のため正しい情報抽出ができないという記事が減少したためである．

テンプレート集合 単一項目テンプレートにおいてT1とT2の結果を比較してみると，ほとんど結果は悪くなっておらず，「製品種別」にいたっては結果は良くなっている．このことにより，包含関係を用いた削減処理では，テンプレートの質はほとんど変化しないと言える．
また，単一項目テンプレートにおいてT2とT3の結果を比較してみると，全ての項目で結果は悪くなっており，特に「販売元」16%も低くなっている．この様に結果が悪くなったのは，次の様な理由が考えられる．

- 統合したパターンの並びを考慮しなかったため，テンプレートとのマッチングによって本来抽出すべき情報が抽出出来なかったため
- 新たにパターンを統合して出来たテンプレートに重みを付けなおしたことによって，解候補の選択の精度が悪くなったため

抽出項目 「製品種別」の結果が他の4項目に比べて極端に低い．これは，「製品種別」を表す抽出情報の前後に現れる文字種は他の抽出項目に比べてかなり多いため(表1)，その結果「製品種別」に関するテンプレートの数が多くなり，パターンマッチングにより多数の解候補が得られ，そのため解候補の選択がうまくいかないためである．
それに対して，「販売元」と「発売日」を比較すると，対応する抽出情報の前後に現れる文字種は「発売日」の方が多いのにもかかわらず，抽出結果は「販売元」より高い結果となっている．これは，「発売日」はテンプレートのマッチングで抽出された情報に対して，「発売日」固有の表現を用いた制約を用いて解候補を制限しているため，このような結果となっている．「価格」の抽出精度が高い結果となっているのも同様の理由である．

6 おわりに

テンプレートを用いた新聞記事からの製品情報抽出において，単一項目テンプレートを用いることによって，テンプレート数を増加させず，再現率を向上させることができることを示した．また，抽出精度をほとんど落とさずにテンプレートを大幅に削減する方法を示した．

今回は，1記事中に複数製品を紹介している記事については考慮していない．今後，これらの記事を扱えるように，処理方式を改良できれば，さらなる抽出精度の向上が期待できる．

さらに，複数の解の中から，適切なものを選択する方法については，さらなる検討が必要である．「製品種別」の場合，この解の選択法を改良することにより抽出精度が大幅に改善されることが期待できる．

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] 井出 裕二, 藤吉 誠, 永井 秀利, 中村貞吾, 野村浩郷：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, 情報処理学会研究報告 96-NL-115, pp. 83 – 90, 1996

[2] 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 97-NL-118, pp. 7 – 14, 1997

[3] 柴田 和査：単一項目テンプレートによる新聞記事からの製品情報抽出, 平成8年度九州工業大学卒業論文 (1997)

[4] 江里口 善生, 木谷 強：冨田一般化LRパーザを用いた情報抽出, 情報処理学会論文誌 Vol.38 No.1, pp.44 – 54, 1997

[5] 江里口 善生, 木谷 強：冨田一般化LRパーザを用いた情報抽出, 情報処理学会研究報告 94-NL-102, pp. 9 – 16, 1994

[6] 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

[7] 河合 敦夫, 塚本 雄之, 椎野 努：電子メール文書からの関係情報の自動抽出, 情報処理学会研究報告 94-NL-101, pp. 57 – 64, 1994